IJCSIS Vol. 15 No. 12, December 2017
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2017 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Editorial Board

*It is our great pleasure to present the **December 2017 issue** (Volume 15 Number 12) of the **International Journal of Computer Science and Information Security** (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and technology. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 9800 times and this journal is experiencing steady and healthy growth. Google statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, Academia.edu and EBSCO among others.*

A great journal cannot be made great without a dedicated editorial team of editors and reviewers. On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers for their outstanding efforts to review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status, making sure we deliver high-quality content to our readers in a timely fashion.

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." We would like to thank you, the authors and readers, the content providers and consumers, who have made this journal the best possible.

*For further questions or other suggestions please do not hesitate to contact us at **ijcsiseditor@gmail.com**.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 15, No. 12, December 2017 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	**Dr Riktesh Srivastava [Profile]** Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	**Dr. Jianguo Ding [Profile]** Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran	**Dr. Naseer Alquraishi [Profile]** University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	**Dr. Kai Cong [Profile]** Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	**Dr. Omar A. Alzubi [Profile]** Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	**Dr. Jorge A. Ruiz-Vanoye [Profile]** Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	**Prof. Ning Xu,** Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	**Dr . Bilal Alatas [Profile]** Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	**Dr. Ioannis V. Koskosas,** University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	**Dr Venu Kuthadi [Profile]** University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	**Dr. Zhihan lv [Profile]** Chinese Academy of Science, China
Dr. Vahid Esmaeelzadeh [Profile] Iran University of Science and Technology	**Prof. Ghulam Qasim [Profile]** University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	**Prof. Dr. Maqbool Uddin Shaikh [Profile]** Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	**Dr. Musa Peker [Profile]** Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	**Dr. Wencan Luo [Profile]** University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	**Dr. Ijaz Ali Shoukat [Profile]** King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	**Dr. Yilun Shang [Profile]** Tongji University, Shanghai, China
Dr. Sedat Akleylek [Profile] Ondokuz Mayis University, Turkey	**Dr. Sachin Kumar [Profile]** Indian Institute of Technology (IIT) Roorkee

ISSN 1947 5500

TABLE OF CONTENTS

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

6. PaperID 30111706: Robotics and Artificial Intelligence: Differences and Similarities (pp. 26-28)

(1) Omankwu, Obinnaya Chinecherem; (2) Nwagu, Chikezie Kenneth; (3) Inyiama, Hycient
(1) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(2) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

7. PaperID 30111707: Faculty Assignment and Timetabling using Optimisation (pp. 29-35)

MA Nang Laik, Senior Lecturer, School of Business, Singapore University of Social Sciences, Singapore
SEN Prabir, Director Research, STATGRAF Research, Vancouver, Canada

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

8. PaperID 30111709: A Novel and Robust Retroareolar Region Detection in Mammography Images (pp. 36-39)

Dr. P. Pedda Sadhu Naik, Professor & HOD, Dr. Samuel George Institute of Engineering & Technology, Markapur, AP.
Dr T. Venugopal, Professor & HOD, JNTUH College of Engineering, Jagityala, Telangana

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

9. PaperID 30111712: Improving DBMS Security through the use of a Checksum Technique (pp. 40-48)

*Naim Ajlouni *, Firas Ajlouni **, Alaa Ali Hameed ****
*Istanbul Aydin University * ****
*Lancashire College of Further Education ***

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

10. PaperID 30111713: Coupling GIS-based MCA and AHP techniques for Hospital Site Selection (pp. 49-56)

Aseel B. Kmail, Jacqueleen Jubran, Walid Sabbah
Geographic Information System Department, AAUJ, Jenin, Palestine

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

11. PaperID 30111714: Big and Connected Data Analysis with Graph and Relational Databases Using Collaborative Filtering Technique (pp. 57-65)

Joseph George Davis, Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
Mr Joseph K. Panford, Senior Lecturer, Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

(1) Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
(2) Computer Science Department, Michael Okpara University of Agriculture Umudike Umuahia, Abia State, Nigeria.
(3) Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

23. PaperID 30111741: Credit Card Fraud Detection Using Neural Network (pp. 176-180)

(1) Nwagu, Chikezie Kenneth; (2) Omankwu, Obinnaya Chinecherem; (3) Inyiama, Hycient
(1) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(2) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

24. PaperID 30111742: Reinforcement Learning a Tool for Filtering Personalized Web Document (pp. 181-185)

(1) Nwagu, Chikezie Kenneth; (2) Omankwu, Obinnaya Chinecherem; (3) Inyiama, Hycient
(1) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(2) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

25. PaperID 30111744: Influence of Distributed Generation on the Protection Scheme of Local Distribution System (pp. 186-190)

Usman Yousaf, Dr. Azzam ul Asar, Alishpa Khattak
Department of Electrical Engineering, CECOS University, Peshawar, Pakistan

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

26. PaperID 30111747: Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm (pp. 191-200)

Abdeljalil EL ABDOULI, Larbi HASSOUNI, Houda ANOUN
RITM Laboratory, CED Engineering Sciences, Ecole Superieure de Technologie, Hassan II University of Casablanca, Morocco

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

27. PaperID 30111749: Using Biometric Watermarking for Video File Protection Based on Chaotic Principle (pp. 201-206)

Saja J. Mohammed, Computer Science Department, College of Computer Science & Mathematics, Mosul University Mosul, IRAQ

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

Dr Narasimha Rao Yamarthi (1), Mr Adinew Belay (2), Mr Abdisa Lechisa (3), P.Narasimha Rao (4).
(1) Professor, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
(2) HOD, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
(3) Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
(4) Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

41. PaperID 30111773: An Efficient Voice Based Person Identification System for Secured Communication (pp. 308-325)

Dr. Shakila Basheer, Lecturer in Information system Department, King Khalid University, ABHA, Saudi Arabia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

42. PaperID 30111766: Comparative Analysis of Modern Methods and Algorithms of Cryptographic Protection of Information (pp. 326-330)

Saleh I. Alomar & Saleh A. Khawatreh
AL-Ahliyya Amman University, Department of Engineering

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

43. PaperID 30111775: SQS: An Ontology Based Framework for Evaluating Service Oriented Software Applications: A case study of E-Governance Services (pp. 331-336)

Ezekiel U, Okike, University of Botswana, Gaborone, Botswana

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

Risk Management Process Analysis for Information & Communication Technology (ICT) Systems: Risk Assessment perspective

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem, Okonkwo, Obikwelu Raphael

[1]Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
Nwaguchikeziekenneth@hotmail.com

[2]Computer Science Department, Michael Okpara University of Agriculture Umudike Umuahia, Abia State, Nigeria.
Saintbeloved@yahoo.com

[3]Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
Oobi2971@yahoo.com

Abstract— ***This Research Paper, explored Risk management from three perspectives: Risk Assessment, Risk Mitigation & Continuous Risk monitor, evaluate and*** Review***. Risk Assessment Perspective highlights vulnerability/threat sources, types of Security, methods used to "break" security. The assessment processes were fully depicted showing its specific steps. It also reflected importance of consistent communication and consultation among the stakeholders as it is key and sustained throughout the entire processes. This work also strengthens the fact that risk is not entirely negative as reveals by its SWOT (Strength, Weakness, Threat and Opportunity) analysis. Hence one can build on the opportunity thereby reducing the associated threat and also improve on the strengths. This invariably reduces and if possible eradicates the weaknesses.***

Keywords—Risk, Risk Management, Risk Assessment, Vulnerability, threat, Opportunity

I. INTRODUCTION

In the face of global ever-dynamic threats and attacks, every Organization is adopting measures to reduce negative risks and utilize positive risks. This ensures that her vision and mission is protected, guarded and fully enhanced. This critical as Organizations make ICT, hub for better support and sustenance of her business. As Organizations automated their processes using data and communication devices, Risk Management plays a very critical role in protecting the organizations information assets, and therefore its mission. Risk Management is every stakeholder`s duty and not only for the technical IT team. Therefore, it should be treated as fundamentally as an essential role of the Management. An effective and efficient risk management process is an important component of a successful ICT security so as to ensure data confidentiality, integrity and high availability. **Vulnerability** is the devices` weakness which can be accidentally triggered or intentionally exploited. While **opportunity** is just positive risk which can be invested upon in order to maximize the use and benefit the devices. Based on this, Organizations are continuously working on reducing the vulnerabilities by minimizing sources of threats and maximizing the opportunities by strengthening the securities during **SWOT (Strength-weakness-Opportunities and threats)** analysis. **Risk management** is the process of identifying risk, assessing risk, and taking steps to reduce risk to an acceptable level, if possible eradicate it completely. The objectives of Risk assessment in this work are to increase the likelihood and impact of positive events, and decrease the likelihood of negative events in devices. However, before now, Risk management is not consciously or transparently carried out for data and communication devices as the "practice of the day" is that the Organization`s perceived final step in the system development life cycle of the devices is always junking of the devices without final proper risk assessment to ensure that no critical piece of information or data can be intentional or accidentally exploited.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

II. VULNERABILITIES/THREAT SOURCES

Risks are continuously evolving. The goal of this step is to identify the potential vulnerability and threat-sources that are applicable to the ICT systems being evaluated. A threat-source is defined as any circumstance or event with the potential to cause harm to an ICT system. The common threat-sources can be natural, human, or environmental:

- Natural Threats: Floods, earthquakes, tornadoes, landslides, avalanches, electrical storms, and other such events.
- Human Threats: Events that are either enabled by or caused by human beings, such as unintentional acts (inadvertent data entry) or deliberate actions (network based attacks, malicious software upload, unauthorized access to confidential information).
- Environmental Threats: Long-term power failure, pollution, chemicals, liquid leakage.

In assessing threat-sources, it is important to consider all potential threat-sources that could cause harm to an ICT System and its processing environment. For example, although the threat statement for an ICT System located in a desert may not include natural flood, because of the low likelihood of such an events occurring, environmental threats such as a bursting pipe can quickly flood a computer room and cause damage to an organization's IT assets and resources.

Humans can be threat-sources through intentional acts, such as deliberate attacks by malicious persons or disgruntled employees, or unintentional acts, such as negligence and errors.

A deliberate attack can be either:

- A malicious attempt to gain unauthorized access to an ICT System (e.g., via password guessing) in order to compromise system and data integrity, availability, or confidentiality or
- A benign, but nonetheless purposeful, attempt to circumvent system security.

Motivation and Threat Actions

According to Siciliano, (2011) Hackers are motivated by a number of factors such as ego, religion, politics, activism etc. Motivation and the resources for carrying out an attack make humans potentially dangerous threat-sources. In addition, reviews of the history of system break-ins; security violation reports; incident reports; and interviews with the system administrators, help desk personnel, and user community during information gathering help identify human threat-sources that have the potential to harm an IT system and its data and that may be a concern where vulnerability exists.

With these information, organizations should be mindful of them and consciously have proofs against them to reduce and/or prevent successful exploits.

Types of Security

Secure communication is when two or more devices are communicating without eavesdropping or interception by a third party. This communication involves sharing of data and information with varying confidentiality and integrity. Among the means to achieve this is:

- Code: This is a means whereby the content and nature of communication is hidden. It is a rule to convert a piece of information and data (for example, a letter, word, phrase or gesture) into another form of representation, not necessarily of same type.
- Encryption: This is also another means whereby the nature and content of communication is hidden. Here, data and communication is rendered hard to read to any unauthorized party. In some highly security-conscious environments, encryption is configured such that it is a basic requirement for connection and communication to be established. No room for opportunistic encryption which is a lower security method to generally increase percentage of generic traffic and this makes the content susceptible to eavesdropping.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Steganography: This is sometimes referred to as "hidden writing" in which data can be hidden within another, mostly innocuous data. In this way, it is difficult to find or remove unless you know how to find it.For example in communication, the hiding of important data such as telephone number in apparently innocuous data (an MP3 music file). A good advantage of this is plausible deniability – unless one can prove that the data is there (which is usually not easy), it is deniable that the file contains any.
- Identity based Networks: Unwanted or malicious behaviour is possible on the web since it is inherently anonymous. Identity based network removes the chance of anonymity as the identity of the sender and recipient are known.
- "Security by Obscurity: Similar to needle in a haystack in which secrecy of design or implementation is used to provide security. Though this is discouraged and not recommended by standard bodies. But stakeholders believe that if the flaws are not known, then attackers will be unlikely to find them. As it is known that attacker`s first step is usually information gathering which is delayed by this.
- Random Traffic: This involves creating random data flow to make the presence of genuine communication harder to detect and traffic analysis less reliable.
- Hard to trace routing methods: This method hides the parties involved in a communication through unauthorized third-party systems or relays.

Methods used to "break" security

- Bugging: This is simply known as covert listening device which involves miniature transmitter and microphone. This enables unauthorized parties to listen to conversation.
- Computers (general): Any security obtained from a computer is limited by the many ways it can be compromised – by hacking, keystroke logging, backdoors or even in extreme cases by monitoring the tiny electrical signals given off by keyboard or monitors to reconstruct what is typed or seen.
- Laser audio Surveillance: Sounds including speech inside rooms can be sensed by bouncing a laser beam off a window of the room where a conversation is held and detecting and decoding the vibrations in the glass caused by the sound waves.
- Spoofing: This is a situation in which one person or program successfully masquerades as another by falsifying data and thereby gaining an illegitimate advantage or access. For example, Caller Id, Email address, IP address etc. can all be spoofed.

III. RISK ASSESSMENT PROCESSES

Risk Management involves three main principal processes: risk assessment, risk Mitigation and, Monitor, evaluate and review. Risk Assessment has many steps (see figure 1) with processes as broadly summarized:

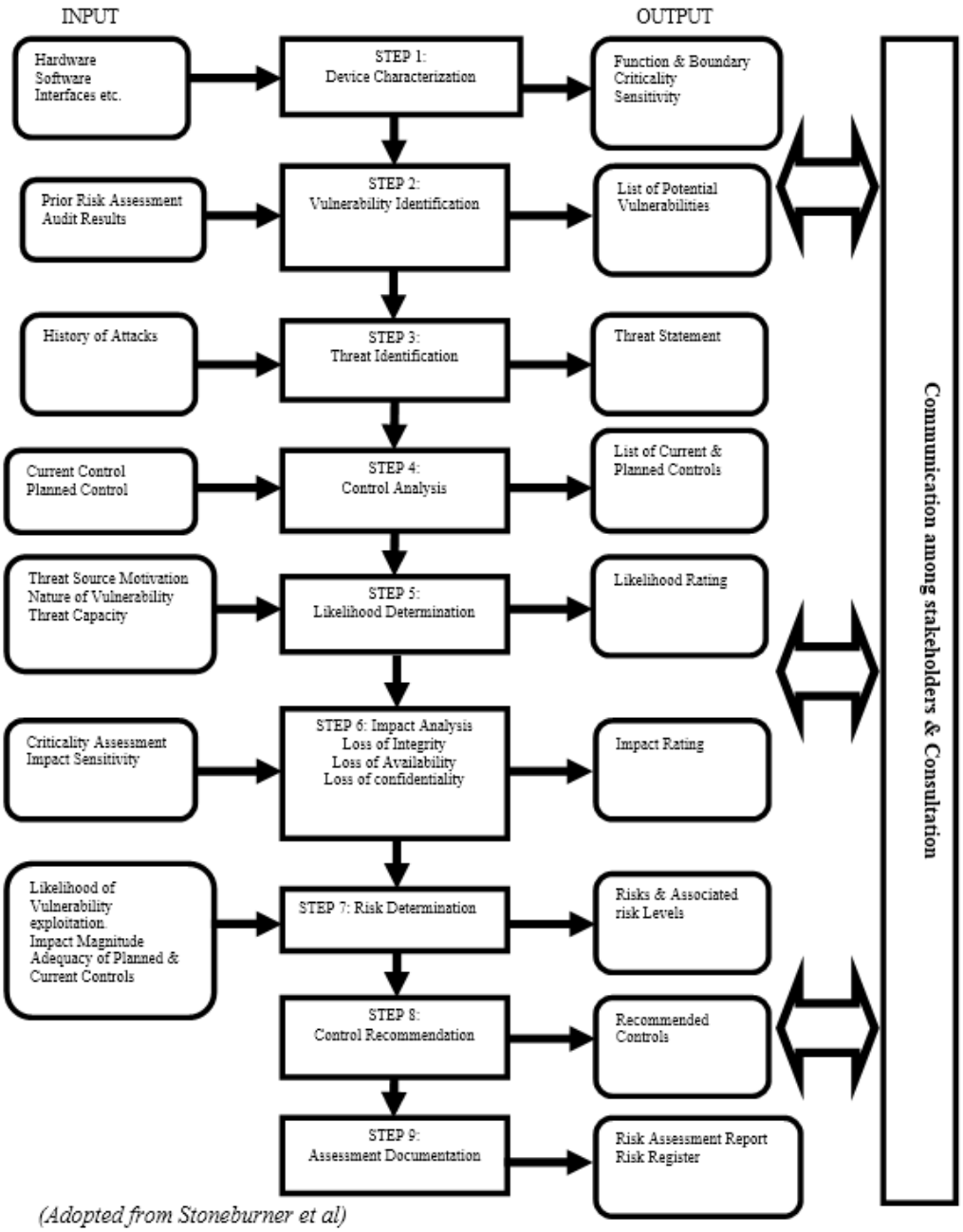
- Risk identification: This allows individuals to identify risks so that the stakeholders will be in the know of potential problems inherent in the devices. It is pertinent to start this stage as early as possible and should be repeated frequently.
- Risk analysis and Priority: Risk analysis transforms the estimates or data about specific risks that developed during risk identification into a consistent form that can be used to make decisions around prioritization. Risk prioritization enables operations to commit resources to manage the most important risks.
- Risk register (Statements) integration: This is the result of risk assessment process. It is a document which contains lists of identified risks, root causes of risks, lists of potential responses, risk owners, symptoms and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

warning signs, relative rating or priority list. Risk for additional analysis and responses, and a watch list which is a list of low-priority risk within the risk register.

- Consistent Communication and Consultation: There is steady communication among stakeholders within the organization as everyone is practically involved. In addition to this, the stakeholders can consult the manufacturer of the device through any of the appropriate channels such as through their Representative or customer voices. This ensures speedy and reliable Response.

Figure I : RISK ASSESSMENT PROCESSES CYCLE

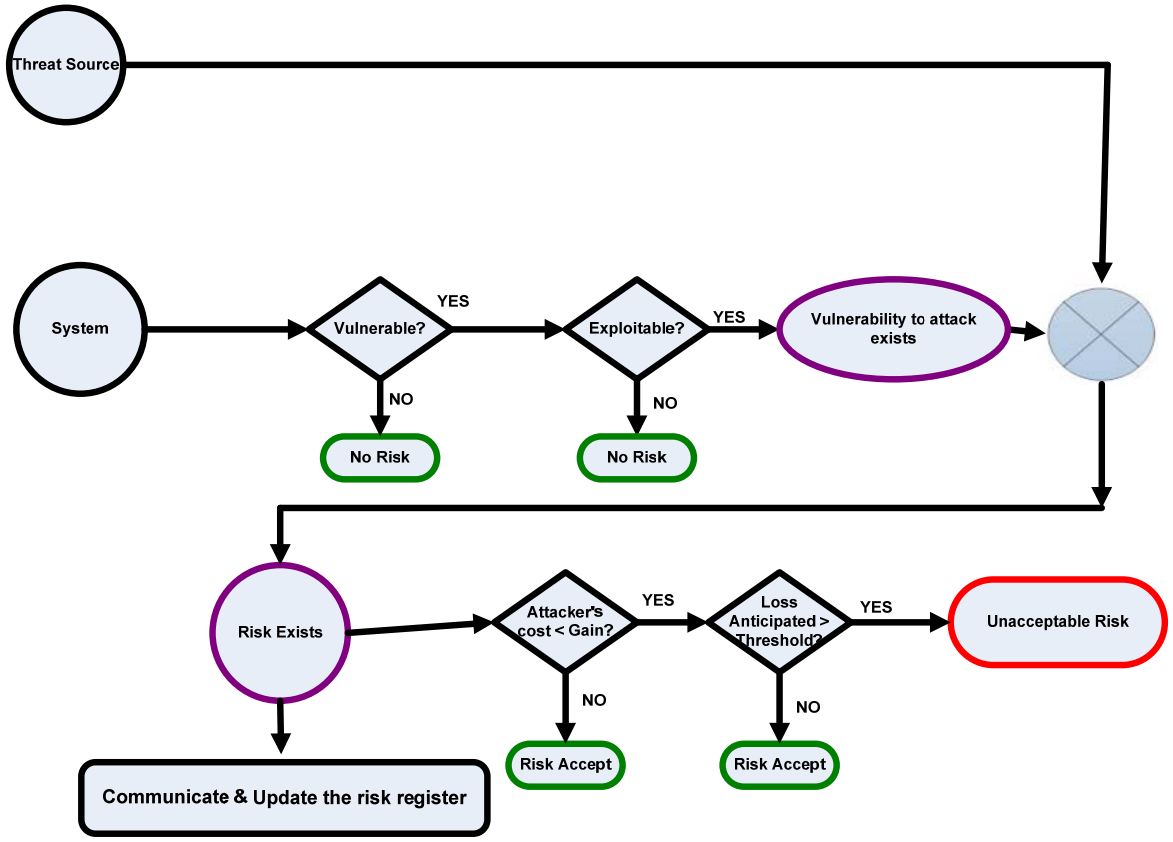


(Adopted from Stoneburner et al)

The processes are followed up with consistent updates and awareness campaigns among stakeholders as new challenges, discoveries and prospects arise.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure II: SIMPLIFY OVERALL RISK ASSESSMENT PROCESS FLOWCHAT FOR ICT SYSTEM



(Adopted from Stoneburner et al)

Principal stakeholders, knowing the potential risks and recommended controls, may ask, "When and under what circumstances should I take action? When shall I implement these controls to mitigate the risk and protect our organization?" Below system architecture is the answer. The system architecture is further articulated in the following rules of thumb, which provide guidance on actions to mitigate risks from intentional human threats:

- When vulnerability (or flaw, weakness) exists ➞ implement initial risk assessment to reduce the likelihood of a vulnerability's being exercised.
- When vulnerability can be exercised ➞ apply layered protections, architectural designs, and administrative controls to minimize the risk or prevent the occurrence.
- When the attacker's cost is less than the potential gain ➞ apply protections to decrease an attacker's motivation by increasing the attacker's cost (e.g., use of system controls such as limiting what a system user can access and do can significantly reduce an attacker's gain), communicate and update the risk registers.
- When loss is too great ➞ apply design principles, layered architectural designs, and technical and nontechnical protections to limit the extent of the attack, thereby reducing the potential for loss, update the risk register and communicate.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. CONCLUSIONS

One of the main aims of Risk Management process analysis for ICT systems: Risk Assessment perspective is to ensure that the devices are deployed with best of security measures in place, which makes the organization to be fully proactive rather than reactive as it is today with many organizations. This will, invariably increase stakeholders ` risk appetite for positive risks and of course, establish a careful risk threshold for negative risks. This further ensures that cost of attack from a potential intentional attacker is far higher than the anticipated gain which would likely discourage the attackers. A successful attack has high currency impact, loss of customer confidence and negative business reputation. It assists management to make well-informed risk management decisions to justify huge capital expenditures that are part of an ICT budget and also in authorizing the ICT devices on the basis of the supporting documentation resulting from the performance of risk management. With the flowchart, stakeholders are convinced of what to do, hence proactively take the right steps/decision to protect the organization and ensure optimal utilization of the ICT systems.

It is worthy of note here that the process is a continuous one in order to get optimal throughput from the devices with little or no down time as a result of attack, vulnerability or negative risks .Thereby increasing the opportunities which the devices can offer. Therefore Risk management continues even at the final stage of systems development Life cycle which is disposal of the devices. It is pertinent to carry out risk assessment at the disposal stage to ensure sensitive data or information are not left out such as vital configurations which may include plain-text passwords, administrator credentials etc.

V REFERENCES

Stoneburner G.,Goguen A. and Feringa A.(2001,July). Risk Management Guide for information Systems. Retrieved September 4,2014 from http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf.

Radack S.(Ed.).Security for wireless network and Devices. Retrieved August 20,2014 from http://www.itl.nist.gov/lab/bulletns/bltnmar03.htm

Cooney M.(2012,September 21).10 common mobile Security problems to attack. Retrieved August 22,2014 from http://www.pcworld.com/article/2010278/10-common-mobile-security-problems-to-attack.html

A Guide to the Project Management Body of Knowledge(PMBOK) Fifth Edition.

Alshboul A.(2010).Information Systems Security measures and countermeasures: Protecting Organizational Assets from malicious Attacks,3,Article 486878.Retrieved September 10,2014 from http://www.ibimapublishing.com/journals/CIBIMA/2010/486878/486878.pdf

Secure Communication. Retrieved September 9,2014 from http://en.wikipedia.org/wiki/Secure_communication.

http://www.labcompliance.com/tutorial/risk/default.aspx?sm=d_a. Accessed September 9,2014

http://www.iip.kit.edu/english/1680.php Accessed September 7,2014

Risk Management Process Overview. Accessed September 3,2014 from http://technet.microsoft.com/en-us/library/cc535304.aspx

Schneier, B. (2008). Security through Obscurity.

Siciliano, R. S. (2011). Seven Types of Hacker Motivations. Infosec Island

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Evolutionary Perspective of Robotics

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem, and Inyiama, Hycient

[1] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
Nwaguchikeziekenneth@hotmail.com

[2] Computer Science Department, Michael Okpara University of Agriculture, Umudike
Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka
Anambra State, Nigeria.

ABSTRACT

The evolution of application fields and their sophistication have influenced research topics in the robotics community. This evolution has been dominated by human necessities. In the early 1960s, the industrial revolution put industrial robots in the factory to release the human operator from risky and harmful tasks. The later incorporation of industrial robots into other types of production processes added new requirements that called for more flexibility and intelligence in industrial robots.

Keywords: Robotics, artificial intelligence, algorithm.

.

Introduction

During the last 45 years, robotics research has been aimed at finding solutions to the technical necessities of applied robotics. The evolution of application fields and their sophistication have influenced research topics in the robotics community. This evolution has been dominated by human necessities. In the early 1960s, the industrial revolution put industrial robots in the factory to release the human operator from risky and harmful tasks. The later incorporation of industrial robots into other types of production processes added new requirements that called for more flexibility and intelligence in industrial robots. Currently, the creation of new needs and markets outside the traditional manufacturing robotic market (i.e., cleaning, demining, construction, shipbuilding, agriculture) and the aging world we live in is demanding field and service robots to attend to the new market and to human social needs.

This article addresses the evolution of robotics research in three different areas: robot manipulators, mobile robots, and biologically inspired robots. Although these three areas share some research topics, they differ significantly in most research topics and in their application fields. For this rea-son, they have been treated separately in this survey. The section on robot manipulators includes research on indus-trial robots, medical robots and rehabilitation robots, and briefly surveys other service applications such as refueling, picking and palletizing. When surveying the research in mobile robots we consider terrestrial and underwater vehi-cles. Aerial vehicles are less widespread and for this reason have not been considered. Biologically inspired robots include mainly walking robots and humanoid robots; how-ever, some other biologically inspired underwater systems are briefly mentioned. In spite of the differences between robot manipulators, mobile robots and biologically inspired robots, the three research areas converge in their current and future intended use: field and service robotics. With the modernization of the First World, new services are being demanded that are shifting how we think of robots from the industrial viewpoint to the social and personal viewpoint. Society demands new robots designed to assist and serve the human being, and this harks back to the first origins of the concept of the robot, as transmit-ted by science fiction since the early 1920s: the robot as a human servant (see Figure 1). Also, the creation of new needs and markets outside the traditional market of manufacturing robotics leads to a new concept of robot. A new sector is therefore arising from robotics, a sector with a great future giving service to the human being. Traditional industrial robots and mobile robots are being modified to address this new market. Research has evolved to find solutions to the technical necessities of each stage in the development of service robots.

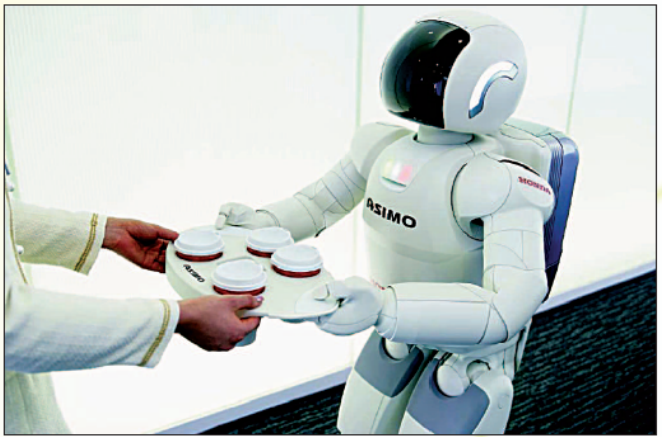


Figure 1. *ASIMO. Photograph courtesy of American Honda Motor Co.*

Robot Manipulators

A robot manipulator, also known as a robot arm, is a serial chain of rigid limbs designed to perform a task with its end-effector. Early designs concentrated on industrial manipulators, to perform tasks such as welding, painting, and palletizing. The evolution of the technical necessities of society and the technological advances achieved have helped the strong growth of new applications in recent years, such as surgery assistance, rehabilitation, automatic refuelling, etc. This section surveys those areas that have received a special, concentrated research effort, namely, industrial robots, medical robots, and rehabilitation robots.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Industrial Robots

It was around 1960 when industrial robots were first introduced in the production process, and until the 1990s industrial robots dominated robotics research. In the beginning, the automotive industry dictated the specifications industrial robots had to meet, mainly due to the industry's market clout and clear technical necessities. These necessities determined which areas of investigation were predominant during that period.

One such area was kinematic calibration, which is a necessary process due to the inaccuracy of kinematic models based on manufacturing parameters. The calibration process is carried out in four stages. The first stage is mathematical modeling, where the Denavit-Hartenberg (DH) method and the product-of-exponential (POE) formulation lead the large family of methods. A detailed discussion of the fundamentals of kinematic modeling can be found in the literature [1]. The gap between the theoretical model and the real model is found in the second stage by direct measurement through sensors. Thus, the true position of the robot's end effector is determined, and by means of optimization techniques, the parameters that vary from their nominal values are identified in the third stage. Last, implementation in the robot is the process of incorporating the improved kinematic model. This process will depend on the complexity of the machine, and iterative methods will have to be employed in the most complex cases. Research in robot calibration remains an open issue, and new methods that reduce the computational complexity of the calibration process are still being proposed [2], [3].

Another important research topic is motion planning, wherein subgoals are calculated to control the completion of the robot's task. In the literature there are two types of algorithms, implicit methods and explicit methods. Implicit methods specify the desired dynamic behavior of the robot. One implicit scheme that is attractive from the computational point of view is the potential field algorithm [4]. One disadvantage of this approach is that local minima of the potential field function can trap the robot far from its goal. Explicit methods pro-vide the trajectory of the robot between the initial and final goal. Discrete explicit methods focus on finding discrete collision-free configurations between the start and goal configurations. These methods consist mainly of two classes of algorithms, the family of road-map methods that include the visibility graph, the Voronoi diagram, the free-way method and the Roadmap algorithm [5], and the cell-decomposition methods [6]. Continuous explicit methods, on the other hand, consist in basically open-loop control laws. One important family of methods is based on optimal-control strategies [7], whose main disadvantages are their computational cost and dependence on the accuracy of the robot's dynamic model.

Besides planning robot motion, control laws that assure the execution of the plan are required in order to accomplish the robot's task. Thus, one fundamental research topic focuses on control techniques. A robot manipulator is a nonlinear, multi-variable system and a wide spectrum of control techniques can be experimented here, ranging from the simpler proportional derivative (PD) and proportional integral derivative (PID) control to the computed-torque method [8], and the more sophisticated adaptive control [9] whose details are out of the scope of this survey.

Typical industrial robots are designed to manipulate objects and interact with their environment, mainly during tasks such as polishing, milling, assembling, etc. In the control of the interaction between manipulator and environment, the contact force at the manipulator's end effector is regulated. There are diverse schemes of active force control, such as stiffness control, compliant control, impedance control, explicit force control and hybrid force/position control.

The first three schemes belong to the category of indirect force control, which achieves force control via motion control, while the last two methods perform direct force control by means of explicit closure of the force-feedback loop. Readers who wish to study this subject in detail will find an interesting account in [10].

An attractive alternative for implementing force-control laws is the use of passive mechanical devices so that the trajectory of the robot is modified by interaction forces due to the robot's own accommodation. An important example of passive force control is the remote center of compliance (RCC) sys-tem patented by Watson in 1978 [11] for peg-in-hole assembly. Passive force control is simpler than active force control laws but has disadvantages, such as lacking flexibility and being unable to avoid the appearance of high contact forces.

Figure 2. *Robots in the food industry.*

As 1990 began, new application areas for industrial robots arose that imposed new specifications, with flexibility as the principal characteristic. The new industries that introduced industrial robots in their productive process were the food and pharmacy industries (see Figure 2). Postal services too looked for robotic systems to automate their logistics. The main requirement was the capacity to accommodate variations in product, size, shape, rigidity (in the case of foods), etc. The ability to self-adapt to the product and the environment became the issue in the following lines of investigation in the area of industrial robotics. The main line of research now is aimed at equipping the control system with sufficient intelligence and problem-solving capability. This is obtained by resorting to artificial-intelligence techniques. Different artificial intelligence (AI) techniques are used to provide the robot with intelligence and flexibility so it can operate in dynamic environments and in the presence of uncertainty. Those techniques belong to three areas of artificial intelligence: learning, reasoning and problem solving [12]. Among the diverse learning algorithms, inductive learning is the most widely used in robotics, in which the robot learns from preselected examples

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Typical reasoning paradigms in robotics include fuzzy reasoning [14], mostly used in planning under uncertainty, spatial reasoning, and temporal reasoning. The techniques most commonly used in robotics for problem solving are means-end reasoning, heuristic searching, and the blackboard (BB) model.

Another solution to the control of robots in dynamic or unknown environments consists of introducing the operator in the control loop, such that the robot is remotely operated. The success of a teleoperation system relies on the correct feedback of the robot interaction with the environment, which can be visual, tactile or force reflection. The greatest disadvantage that teleoperated systems involve are transmission delays when the distance between the operator and the robot is significant, like in space teleoperation or over the Internet. Some research has explored solutions to this problem, such as interposing a virtual robot in charge of environment feedback, but this procedure is only valid if the robot works in structured environments. Another solution is teleprogramming, in which the operator sends high-level commands and the robot carries out the task in closed-loop control. Recently, considerable attention has been devoted to Internet-based teleoperation, in which the transmission delay is variable. For direct force feedback, wave-variable-based approaches have been used extensively, and they have been further extended to include estimation and prediction of the delay. A comprehensive survey can be found in [15].

With the rapid modernization of the First World, new types of services are being required to maintain a certain quality of life. A new, promising robotics sector is arising to serve the human being. Traditional industrial robots are being modified to respond to this new market, yielding surgery robots, refueling robots, picking and palletising robots, feeding robots, rehabilitation robots, etc. Two of the most relevant service applications of robot manipulators are in the field of medical robots and rehabilitation robots that are catching the interest of researchers all over the world. In the following sub-sections, we will summarize research topics in medical robotics and rehabilitation robotics.

Medical Robots

In recent years, the field of medicine has been also invaded by robots, not to replace qualified personnel such as doctors and nurses, but to assist them in routine work and precision tasks. Medical robotics is a promising field that really took off in the 1990s. Since then, a wide variety of medical applications have emerged: laboratory robots, telesurgery, surgical training, remote surgery, telemedicine and teleconsultation, rehabilitation, help for the deaf and the blind, and hospital robots. Medical robots assist in operations on heart-attack victims and make possible the millimeter-fine adjustment of prostheses. There are, however, many challenges in the widespread imple-mentation of robotics in the medical field, mainly due to issues such as safety, precision, cost and reluctance to accept this technology.

Medical robots may be classified in many ways: by manipulator design (e.g., kinematics, actuation); by level of autonomy (e.g., preprogrammed versus teleoperation versus constrained cooperative control); by targeted anatomy or technique (e.g., cardiac, intravascular, percutaneous, laparoscopic, micro-surgical); by intended operating environment [e.g., in-scanner, conventional operating room (OR)], etc. Research remains open in the field of surgical robotics, where extensive effort has been invested and results are impressive. Some of the key technical barriers include safety [16], where some of the basic principles at issue are redundancy, avoiding unnecessary speed or power in actuators, rigorous design analysis and multiple emergency stop and checkpoint/restart facilities. Medical human-machine interfaces are another key issue that draws upon essentially the same technologies as other application domains.

Surgeons rely on vision as their dominant source of feedback; however, due to the limited resolution of current-generation video cameras, there is interest in optical overlay methods, in which graphic information is superimposed on the surgeon's field of view to improve the information provided [17]. As surgeons frequently have their hands busy, there has been also interest in using voice as an interface. Force and haptic feedback is another powerful interface for telesurgery applications [18]. Much of the past and present work on telesurgery involves the use of master-slave manipulator systems [19], [20]. These systems have the ability to feed forces back to the surgeon through the master manipulator, although slaves' limitations in sensing tool-to-tissue forces can some-what reduce this ability.

The field of medical robotics is expanding rapidly and results are impressive as a large number of commercial devices are being used in hospitals. However, societal barriers have to be overcome and significant engineering research effort is required before medical robots have wide-spread impact on health care.

Rehabilitation Robots

Activity in the field of rehabilitation robotics began in the 1960s [21] and has slowly evolved through the years to a point where the first commercially successful products are now available. Today, the concept of "rehabilitation robot" may include a wide array of mechatronic devices ranging from artificial limbs to robots for supporting rehabilitation therapy or for providing personal assistance in hospital and residential sites. Examples include robots for neuro-rehabilitation [22], power-augmentation orthosis [23], rehabilitative orthosis, etc. The field of rehabilitation robotics is less developed than that of industrial robotics. Many assistive robotic systems have featured an industrial robot arm for reasons of economy and availability [24]. However, the specifications for robots in these two application areas are very different. The differences arise from the involvement of the user in rehabilitation applications. Industrial robots are typically powerful and rigid to provide speed and accuracy. They operate autonomously and, for reasons of safety, no human interaction is permitted.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Rehabilitation robots must operate more slowly and be more compliant to facilitate safe user interaction. Thus, rehabilitation robotics is more akin to service robotics, which integrates humans and robots in the same task. It requires safety and special attention must be paid to human-machine interfaces that have to be adapted for disabled or nonskilled people operating a specific programming device. It is also recognized that there is a need for research and development in robotics to focus on developing more flexible systems for use in unstructured environments. The leading developments of this type in rehabilitation robotics concern, among other topics, mechanical design (including mobility and end-effectors), programming, control and man machine interfaces [25]. Subsection “Humanoid Robots” of this article expands on new research into human-robot interaction.

Mobile Robots

The term mobile robot describes a robotic system able to carry out tasks in different places and consisting of a platform moved by locomotive elements. The choice of the locomotive system depends firstly on the environment in which the robot will operate. This can be aerial, aquatic or terrestrial In the aquatic and aerial environments, the locomotive systems are usually propellers or screws, although at the seabed legs are also used. The choice of the locomotive system on earth is more complicated due to the variety of terrestrial environments. Wheels, tracks, and legs are typical terrestrial locomotive elements.

Mobility provides robots with enhanced operating capacity and opens up new areas of investigation. Some such areas are common to all mobile robots, like the navigation problem, whereas others deal more specifically with a certain locomo-tion system, like the walking gait.

Practically by the time industrial robots were introduced in the production process, mobile robots were installed in the factory. This was around 1968, and the robots were mainly automated guided vehicles (AGVs), vehicles transporting tools and following a predefined trajectory. Nevertheless, the research in this area deals now with autonomous indoor and outdoor navigation. Autonomous mobile-robot navigation consists of four stages: perception of the environment, self-localization, motion planning and motion generation.

In structured environments, the perception process allows maps or models of the world to be generated that are used for robot localization and motion planning. In unstructured or dynamic environments, however, the robot has to learn how to navigate. Navigation is, therefore, one of the main applications of artificial intelligence to robotics, where learning, reasoning and problem solving come together. The main research in mobile robotics is focusing on robot localization and map generation.

Conclusion

Since the introduction of industrial robots in the automotive industry, robotics research has evolved over time towards the development of robotic systems to help the human in dangerous, risky or unpleasant tasks. As the complexity of tasks has increased, flexibility has been demanded in industrial robots, and robotics research has veered towards adaptive and intelligent systems.

Since 1995, robotics research has entered the field- and service-robotics world, where we can find manipulators, mobile robots and animal-like robots with great perspectives of development and increasing research interest. Surgical robots have been the first successes, and recently different areas in medical-and rehabilitation-robotics applications have arisen. Other examples can be found in the fields of home cleaning, refueling and museum exhibitions, to name just a few areas.

Service-robotics research is also aimed at providing a comfortable, easy life for the human being in an aging world. The United Nations Economic Commission for Europe (UNECE) forecasts strong growth of professional robots in application areas such as humanoid robots, field robots, underwater systems and mobile robot platforms for multiple use in the period of 2005–2008 [86]. The UNECE also forecasts a tremendous rise in personal robots in the next few years. Robotics research has to make a great effort to solve in very few years the challenges of this new field of research, which will be largely determined by interaction between humans and robots. Figure 10 summarizes the evolution of robotics research over the last 50 years.

It is a fact that, during the last decade, the activity in conferences and expositions all over the world has reflected low activity in industrial manipulators and huge activity in other areas related with manipulation in unstructured environments and mobility, including wheeled, flying, underwater, legged and humanoid robots. Maybe the key is that new challenges in manipulation in factories require less research now because factory needs lie in the field of traditional engineering.

With these premises we can conclude: Yes, definitely robotics research is moving from industrial to field and service applications, and most robotics researchers are enthu-siastic about this broad, exciting field. One development that is very representative of the way the field is evolving is the controversy set off by Prof. Engelberger, the creator of the first robotics company, at the 2005 International Robot Exhibition in Tokyo, Japan, when he commented on the needless research by both Japanese companies and scientific institutions for developing toy-like animal and humanoid robots for very doubtful use. Engelberger thus gained many detractors, who have rapidly argued back that these kinds of robots are a necessary step in the evolution towards real robots capable of helping disabled persons, performing dan-gerous work and moving in hazardous places.

Other defenders of the development of human-like per-sonal robots advocate the importance of aiming at such chal-lenging tasks because of the technology that can be developed, which would prove very important from the commercial point of view in other industrial activities.

Maybe behind all the arguments there still lies the human dream of the universal robot—a single device that can per-form any task. Nothing better for that than a device resembling—what else?—a human being. So, let our imagination flies into the world of service robotics, but, please, do not for-get to keep an eye on traditional industrial manipulators.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

References

[1] J.J. Craig, *Introduction to Robotics*. Reading, MA: Addison-Wesley, 2nd ed., 1989.

[2] L. Zhenyu, C. Yinglin, and Q. Daokui, "Research on robot calibra-tion," *Robot*, vol. 24, no. 5, pp. 447–450, 2002.

[3] S. Lei, L. Jingtai, S. Weiwei, W. Shuihua, and H. Xingbo, "Geometry-based robot calibration method," in *Proc. IEEE Int. Conf. Robotics and Automation*, pp. 1907–1912, 2004.

[4] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *Int. J. Robot. Res.*, vol. 5, no. 1, pp. 90–98, 1986.

[5] J.F. Canny, *The Complexity of Robot Motion Planning*, Cambridge, MA: MIT Press, 1988.

[6] J.T. Schwartz and M. Sharir, "On the 'piano movers' problem: 1. The case of two-dimensional rigid polygonal body moving amidst polygo-nal barriers," *Commun. Pure Appl. Math.*, vol. 36, pp. 345–398, 1983.

[7] J. Bobrow, S. Dubowsky, and J. Gibson, "Time-optimal control of robotic manipulators along specified paths," *Int. J. Robot. Rese.*, vol. 4, no. 3, pp. 3–17, 1985.

[8] J.J.E. Slotine and W. Li, *Applied Nonlinear Control*. Upper Sadlle Rikver, NJ: Prentice-Hall, 1991.

[9] K.J. Aström and K.B. Wittenmark, *Adaptive Control*. Reading, MA: Addison-Wesley, 1989.

[10] B. Siciliano and L. Villani, *Robot Force Control*. Norwelll, MA: Kluwer, 1999.

[11] P.C. Watson, "Remote center compliance system," U.S. Patent No. 4098001, Jul. 1978.

[12] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* 2nd ed. Uppper Saddle River, NJ: Prentice Hall, 2003.

[13] R.S. Michalski, J.C. Carbonell, and T.M. Mitchell, *Machine Learning*, Palo Alto, CA: Tioga, 1983.

[14] S.G. Tzafestas, "Fuzzy systems and fuzzy expert control: An overview," *Knowledge Eng. Rev.*, vol. 9, no. 3, pp. 229–268, 1994.

[15] G. Niemeyer and J.J.E. Slotine, "Telemanipulation with time delays," *Int. J. Robot. Res.*, vol. 23, no. 9, pp. 873–890, 2004.

[16] B. Davies, *A Discussion of Safety Issues for Medical Robots*, In R. Taylor, S. Lavallee, G. Burdea, and R. Moesges, Eds. *Computer-Integrated Surgery*. Cambridge, MA: MIT Press, pp. 287–296, 1996.

[17] M. Blackwell, C. Nikou, A. DiGioia, and T. Kanade, "An image over-lay system for medical data visualization," *Med. Image Anal.*, vol. 4, pp. 67–72, 2000.

[18] R. Kumar, P. Berkelman, Gupta, A. Barnes, P.S. Jensen, L.L. Whitcomb, and R.H. Taylor, "Preliminary experiments in cooperative human/robot force control for robot assisted microsurgical manipula-tion," in *Proc. IEEE Int. Conf. Robotics Automation*, pp. 610–617, 2000, San Francisco, CA.

[19] G.H. Ballantyne and F. Moll, "The da vinci telerobotic surgical system: The virtual operative field and telepresence surgery," *Surg. Clin. North Amer.*, vol. 86, no. 6, pp. 1293–1304, 2003.

[20] J. Marescaux and F. Rubino, "The zeus robotic system: Experimental and clinical applications," *Surg. Clin. North Amer.*, vol. 86, no. 6, pp. 1305–1315, 2003.

[21] Y. Kim and A.M. Cook, *Manipulation and Mobility Aids*, In J.G.Webster et al, Eds. *Electronic Devices for Rehabilitation*. London, U.K.: Chapman and Hall, 1985.

[22] H.I. Krebs, B.T. Volpe, M.L. Aisen, and N. Hogan, "Increasing pro-ductivity and quality of care: Robot-aided neuro-rehabilitation," *J. Rehab. Res. Devel.*, vol. 37, no. 6, pp. 639–652, 2000.

[23] K. Kiguchi and T. Fukuda, "A 3DOF exoskeleton for upper-limb motion assist—Consideration of the effect of bi-articular muscles," in *Proc. IEEE Int. Conf. Robotics Automation*, New Orleans, LA, pp. 2424–2429, 2004.

[24] L.J. Leifer, "Rehabilitative robotics, the stanford robotic aid," in *Proc. WESCON*, San Francisco, CA, 1981, pp. 4–15.

[25] G. Bolmsjö, H. Neveryd, and H. Eftring, "Robotics in rehabilitation," *IEEE Trans. Rehab. Eng.*, vol. 3, no. 1 pp. 77–83, Jan. 1995.

[26] J. Leonard and H. Durrant-White, *Directed Sonar Sensing for Mobile Robot Navigation*. Norwell, MA: Kluwer, 1992.

[27] R. Simmons and S. Koenig, "Probabilistic robot navigation in partially observable environments," in *Proc. Int. Joint Conf. Artificial Intelligence*, pp. 1080–1087, 1995.

[28] W. Burgard, D. Fox, D. Hennig, and T. Schmidt, "Estimating the absolute position of a mobile robot using position probability grids," in *Proc. AAAI National Conf. Artificial Intelligence*, pp. 896–901, 1996.

[29] S. Thrun, D. Fox, F. Dellaert, and W. Burgard, "Particle filters for mobile robot localization," A. Doucet, N. de Freitas and N. Gordon, eds. in *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, New York, 2001.

[30] A. Elfes, "Sonar-based real-world mapping and navigation," *IEEE J. Robot. Automation*, vol. 3, no. 2 pp. 249–265, Mar. 1987.

[31] M.J. Matarié, "A distributed model for mobile robot environment-learning and navigation," M.S. thesis, Mass. Inst. Technol., Cambridge, MA, Jan. 1990, MIT AI Lab. Tech. Rep. AITR-1228, May 1990.

[32] H.P. Moravec, "Sensor fusion in certainty grids for mobile robots," *AI Mag.*, vol. 9, no. 2, pp. 61–74, 1988.

[33] R. Chatila and J.-P. Laumond, "Position referencing and consistent world modeling for mobile robots," in *Proc. IEEE Int. Conf. Robotics and Automation*, 1985, pp. 138–145.

[34] J. Leonard, H.F. Durrant-White, and I.J. Cox, "Dynamic map building for an autonomous mobile robot," *Int. J. Robot. Res.*, vol. 11, no. 4, pp. 89–96, 1992.

[35] J.A. Castellanos and J.D. Tardós, *Mobile Robot Localization and Map Building: A Multisensor Fusion Approach*, Norwell, MA: Kluwer, 2000.

[36] J.E. Guivant and E. Nebot, "Optimization of the simultaneous local-ization and map-building algorithm for real-time implementation," *IEEE Trans. Robot. Automat.*, vol. 17, no. 3, pp. 242–257, May 2001.

[37] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proc. AAAI National Conf. Artificial Intelligence*, Edmonton, Canada, 2002, AAAI, pp. 593–598.

[38] F. Lu and E. Milios, "Globally consistent range scan alignment for environment mapping," *Auton. Robots*, vol. 4, pp. 333–349, 1997.

[39] Y. Liu and S. Thrun, "Results for outdoor-slam using sparse extended information filters," in *Proc. IEEE Int. Conf. Robotics Automation*, 2003, 1227–1233.

[40] F. Dellaert, S.M. Seitz, C. Thorpe, and S. Thrun, "EM, MCMC, and chain flipping for structure from motion with unknown correspondence," *Mach. Learn.*, vol. 50, no. 1–2, pp. 45–71, 2003.

[41] R. Biswas, B. Limketkai, S. Sanner and S. Thrun, "Towards object mapping in non-stationary environments with mobile robots," in *Proc. Int. Conf. Intelligent Robots Systems*, pp. 1014–1019, 2002.

[42] D. Wolf and G.S. Sukhatme, "Online simultaneous localization and mapping in dynamic environments," in *Proc. IEEE Int. Conf. Robotics Automation*, 2004, pp. 1301–1306.

[43] S. Thrun, "Robotic mapping: A survey," Tech. Rep. CMU-CS-02-111, School Comp. Sci., Carnegie Mellon Univ., Pittsburgh, PA 15213, Feb. 2002.

[44] S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz, "MINERVA: A second generation mobile tour-guide robot," in *Proc. IEEE Int. Conf. Robotics Automation*, 1999, pp. 1999–2005.

[45] J. Yuh, "Design and control of autonomous underwater robots: A sur-vey," *Auton. Robots*, vol. 8, pp. 7–24, 2000.

[46] D.A. Read, F.S. Hover, and M.S. Triantafyllou, "Forces on oscillating foils for propulsion and maneuvering," *J. Fluids Structures*, vol. 17, pp. 163–183, 2003.

[47] J. Ayers, "Underwater walking," *Arthropod Struct. Devel.*, vol. 33, no. 3, 347–360, 2004.

[48] G. Georgidas, A. German, A. Hogue, H. Liu, C. Prahacs, A. Ripsman, R. Sim, L.A. Torres, P. Zhang, M. Buehler, G. Dudek, M. Jenkin, and E. Milios, "Aqua: An aquatic walking robot," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots Systems*, 2004, Sendai, Japan, pp. 1731–1737.

[49] R.B. McGhee and A.A. Frank, "On the stability properties of quadruped creeping gaits," *Math. Biosci.*, vol. 3, pp. 331–351, 1968.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[50] P. Gonzalez de Santos, E. Garcia, and J. Estremera, *Quadrupedal Locomotion: An Introduction to the Control of Four-Legged Robots*, London, U.K.: Springer-Verlag, 2006.

[51] S.M. Song and K.J. Waldron, *Machines that walk: The adaptive suspension vehicle*. Cambridge, MA: MIT Press, 1989.

[52] T.A. McMahon, "The role of compliance in mammalian running gaits," *J. Experim. Biol.*, vol. 115, pp. 263–282, 1985.

[53] M.H. Raibert, *Legged Robots That Balance*. Cambridge, MA: MIT Press, 1986.

[54] K. Yoneda, H. Iiyama, and S. Hirose, "Intermitent trot gait of a quadruped walking machine dynamic stability control of an omnidirectional walk," in *Proc. IEEE Int. Conf. Robotics Automation*, 1996, pp. 3002–3007, Atlanta, GA.

[55] H. Kimura and Y. Fukuoka, "Biologically inspired adaptive dynamic walking in outdoor environment using a self-contained quadruped robot: 'tekken2'," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots Systems*, 2004.

[56] S. Hirose and K. Kato, "Quadruped walking robot to perform mine detection and removal task," in *Proc. Int. Conf. Climbing Walking Robots*, 1998, pp. 261–266, Brussels, Belgium.

[57] K. Nonami, Q.J. Huang, D. Komizo, N. Shimoi, and H. Uchida, "Humanitarian mine detection six-legged walking robot," in *Proc. Int. Conf. Climbing Walking Robots*, 2000, pp. 861–868, Madrid, Spain.

[58] P. Gonzalez de Santos, E. Garcia, J. Estremera, and M.A. Armada, "DYLEMA: Using walking robots for landmine detection and location," *Int. J. Syst. Sci.*, vol. 36, no. 9, pp. 545–558, 2005.

[59] P. Gonzalez de Santos, M.A. Armada, and M.A. Jimenez, "Ship build-ing with ROWER," *IEEE Robot. Automat.Mag.*, vol. 7, no. 4, pp. 35–43, Dec. 2000.

[60] R. Molfino, M. Armada, F. Cepolina, and M. Zoppi, "Roboclimber the 3 ton spider," *Ind. Robot: Int. J.*, vol. 32, no. 2, pp. 163–170, 2005.

[61] Q. Huang, K. Yokoi, S. Kajit, K. Kaneko, H. Arai, N. Koyachi, and K. Tanie, "Planning walking patterns for a biped robot," *IEEE Trans. Robot. Automat.*, vol. 17, no. 3, pp. 280–289, May 2001.

[62] C. Azevedo, P. Poignet, and B. Espiau, "Moving horizon control for biped robots without reference trajectory," in *Proc. IEEE Int. Conf. Robotics Automation*, 2002, pp. 2762–2767.

[63] L. Roussel, C. Canudas de Wit, and A. Goswami, "Generation of energy optimal complete gait cycles for biped robots," in *Proc. IEEE Int. Conf. Robotics Automation*, 1998, pp. 2036–2041.

[64] M. Vukobratovic, A.A. Frank, and D. Juricic, "On the stability of biped locomotion," *IEEE Trans. Biomed. Eng.*, vol. BME-17, no. 1, pp. 25–36, Jan. 1970.

[65] J.H. Park, "Impedance control for biped robot locomotion," *IEEE Trans. Robot. Automat.*, vol. 17, no. 6, pp. 870–882, Nov. 2001.

[66] I. de la Guia, P. Staroverov, M. Arbulu, and C. *Balaguer*, "Fast algo-rithm for kinematics problems solving of the low-cost legged robot LEROI," in *Proc. Int. Conf. Climbing Walking Robots*, 2002, pp. 775–782.

[67] T. Takahashi and A. Kawamura, "Posture control for biped robot walk with foot toe and sole," in *Proc. IECON '01*, 2001, pp. 329–334.

[68] Q. Huang, Y. Nakamura, and T. Inamura, "Humanoids walk with feedforward dynamic pattern and feedback sensory reflection," in *Proc. IEEE Int. Conf. Robotics Automation*, 2001, pp. 4220–4225.

[69] P. Sardain and G. Bessonnet, "Forces acting on a biped robot. center of pressureózero moment point," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 34, no. 5, pp. 630–637, Oct. 2004.

[70] K. Hirai, M. Hirose, Y. Haikawa, and T. Takenaka, "The development of Honda humanoid robot," in *Proc. IEEE Int. Conf. Robotics Automation*, May 1998, pp. 1321–1326, Leuven, Belgium.

[71] D.G. Caldwell, N. Tsagarakis, P. Artrit, J. Canderle, S. Davis, and G.A. Medrano-Cerda, "Biomimetic and smart technology principles of humanoid design," in *Proc. IEEE/ASME Int. Conf. Advanced Intelligent Mechatronics*, 2001, pp. 965–970.

[72] R. Caballero, T. Akinfiev, C. Manzano, H. Montes, and M. Armada, "Design of the SMART actuated ROBICAM biped robot," in *Proc. Int. Conf. Climbing Walking Robots*, 2002, pp. 409–416.

[73] M. Gienger, K. Loffler, and F. Pfeiffer, "Towards the design of a biped jogging robot," in *Proc. IEEE Int. Conf. Robotics Automation*, 2001, pp. 4140–4145.

[74] K. Nakadai and H. Tsujino, "Towards new human-humanoid communication: Listening during speaking by using ultrasonic directional speaker," in *Proc. IEEE Int. Conf. Robotics Automation*, 2005, pp. 1495–1500.

[75] P. Nilas, P. Rani, and N. Sarkar, "An innovative high-level human-robot interaction for disabled persons," in *Proc. IEEE Int. Conf. Robotics Automation*, 2004, pp. 2309–2314.

[76] Y. Chen and W.S. Newman, "A human-robot interface based on electrooculography," in *Proc. IEEE Int. Conf. Robotics Automation*, 2004, pp. 243–248.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Knowledge Discovery in Databases (KDD): An Overview

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem, and Inyiama, Hycient

[1] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[2] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

Knowledge Discovery in Databases is the process of searching for hidden knowledge in the massive amounts of data that we are technically capable of generating and storing. Data, in its raw form, is simply a collection of elements, from which little knowledge can be gleaned. With the development of data discovery techniques the value of the data is significantly improved. A variety of methods are available to assist in extracting patterns that when interpreted provide valuable, possibly previously unknown, insight into the stored data. This information can be predictive or descriptive in nature. Data mining, the pattern extraction phase of KDD, can take on many forms, the choice dependent on the desired results. KDD is a multi-step process that facilitates the conversion of data to useful information.Our increased ability to gain information from stored data raises the ethical dilemma of how the information should be treated and safeguarded.

Keywords
Knowledge Discovery Databases, Data Mining, Knowledge Mining

1. INTRODUCTION

The desire and need for information has led to the development of systems and equipment that can generate and collect massive amounts of data. Many fields, especially those involved in decision making, are participants in the information acquisition game. Examples include: finance, banking, retail sales, manufacturing, monitoring and diagnosis, health care, marketing and science data acquisition. Advances in storage capacity and digital data gathering equipment such as scanners, has made it possible to generate massive datasets, sometimes called data warehouses that measure in terabytes. For example, NASA's Earth Observing System is expected to return data at rates of several gigabytes per hour by the end of the century (Way, 1991). Modern scanning equipment record millions of transactions from common daily activities such as supermarket or department store checkout-register sales. The explosion in the number of resources available on the World Wide Web is another challenge for indexing and searching through a continually changing and growing "database."

Our ability to wade through the data and turn it into meaningful information is hampered by the size and complexity of the stored information base. In fact, the shear size of the data makes human analysis untenable in many instances, negating the effort spent in collecting the data. There are several viable options currently being used to assist in weeding out usable information. The information retrieval process using these various tools is referred to as Knowledge Discovery in Databases (KDD).

"The basic task of KDD is to extract knowledge (or information) from lower level data (databases) (Fayyad *et al*, 1995). There are several formal definitions of KDD, all agree that the intent is to harvest information by recognizing patterns in raw data. Let us examine definition proposed by Fayyad, Piatetsky-Shapiro and Smyth, "Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad *et al*, 1995). The goal is to distinguish from unprocessed data, something that may not be obvious but is valuable or enlightening in its discovery. Extraction of knowledge from raw data is accomplished by applying Data Mining methods. KDD has a much broader scope, of which data mining is one step in a multidimensional process.

Knowledge Discovery in Databases Process

Steps in the KDD process are depicted in the following diagram. It is important to note that KDD is not accomplished without human interaction. The selection of a data set and subset requires an understanding of the domain from which the data is to be extracted. For example, a database may contain customer address that would not be pertinent to discovering patterns in the selection of food items at a grocery store. Deleting non-related data elements from the dataset reduces the search space during the data mining phase of KDD. If the dataset can be analyzed using a sampling of the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

data, the sample size and composition are determined during this stage.

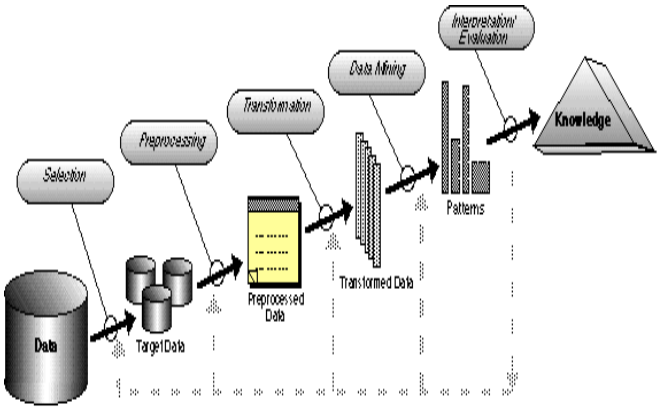


Fig. I: Steps in KDD Process.

Databases are notoriously "noisy" or contain inaccurate or missing data. During the preprocessing stage the data is cleaned. This involves the removal of "outliers" if appropriate; deciding strategies for handling missing data fields; accounting for time sequence information, and applicable normalization of data(Fayyad.1996)

In the transformation phase attempts to limit or reduce the number of data elements that are evaluated while maintaining the validity of the data. During this stage data is organized, converted from one type to another (i.e. changing nominal to numeric) and new or "derived" attributes are defined.

At this point the data is subjected to one or several data mining methods such as classification, regression, or clustering. The data mining component of KDD often involves repeated iterative application of particular data mining methods. "For example, to develop an accurate, symbolic classification model that predicts whether magazine subscribers will renew their subscriptions, a circulation manager might need to first use clustering to segment the subscriber database, and then apply rule induction to automatically create a classification for each desired cluster (Simoudis, 1996). Various data mining methods will be discussed in more detail in following sections.

The final step is the interpretation and documentation of the results from the previous steps. Actions at this stage could consist of returning to a previous step in the KDD process to further refine the acquired knowledge, or translating the knowledge into a form understandable to the user. A commonly used interpretive technique is visualization of the extracted patterns. The results should be critically reviewed and conflicts with previously believed or extracted knowledge resolved.

Understanding and committing to all phases of the data mining process is crucial to its success.

Data Mining Models

A few of the many model functions being incorporated in KDD include:

Classification: mapping or classifying data into one of several predefined classes (Hand, 1981). For example, a bank may establish classes based on debt to income ratio. The classification algorithm determines within which of the two classes an applicant falls and generates a loan decision based on the result.

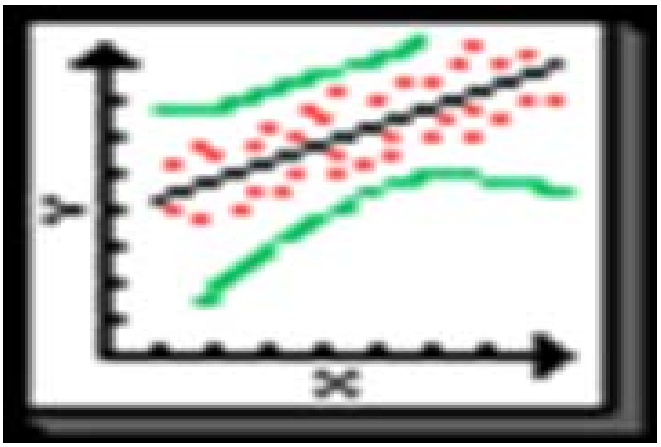


Figure 2: Regression Analysis

Regression: "a learning function which maps a data item to a real-valued prediction variable (Hand, 1981). Comparing a particular instance of an electric bill to a predetermined norm for that same time period and observing deviations from that norm is an example of regression analysis.

Clustering: "maps a data item into one of several categorical classes (or clusters) in which the classes must be determined from the data, unlike classification in which the classes are predefined. Clusters are defined by finding natural groupings of data items based on similarity metrics or probability density models (Fayyad *et al*, 1996). An example of this technique would be grouping patients based on symptoms exhibited. The clusters need not be mutually exclusive.

Summarization: generating a concise description of the data. Routine examples of these techniques include the mean and standard deviation of specific data elements within the dataset.

Dependency modeling: developing a model that shows a how variables are interrelated. An example would be a model showing that electrical usage is highly correlated with the ambient temperature.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Choosing a Data Mining Model

There are no established guidelines to assist in choosing the correct algorithm to apply to a dataset. Typically, the more complex models may fit the data better but may also be more difficult to understand and to fit reliably (Fayyad *et al*, 1995). Successful applications often use simpler models due to the their ease of translation. Each technique tends to lend itself to a particular type problem. Understanding the domain will assist in determining what kind of information is needed from the discovery process thereby narrowing the field of choice. Results can be broken into two general categories; prediction and description. Prediction, as the name infers, attempts to forecast the possible future values of data elements. Prediction is being applied extensively in the area of finance in an attempt to forecast movement in the stock market. Description seeks to discover interpretable patterns in the data. Fraud detection is an application that uses description to identify characteristics of potential fraudulent transactions.

Classification, clustering, summarization and dependency modeling are descriptive models, while regression is predictive.

Current Applications of KDD

Several Knowledge Discovery Applications have been successfully implemented. "SKICAT, a system which automatically detects and classifies sky objects image data resulting from a major astronomical sky survey. SKICAT can outperform astronomers in accurately classifying faint sky objects(Fayyad *et al*, 1995). KDD is being used to flag suspicious activities on two frontiers: Falcon alerts banks of possible fraudulent credit card transactions and the FAIS system being employed by the Financial Crimes Enforcement Network detects financial transactions that may indicate money laundering (Simoudis, 1996). Market Basket Analysis (MBA) has incorporated discovery driven data mining techniques to gain insights about customer behavior. Other applications are being used in the Molecular Biology, Global Climate Change Modeling and other concentrations where the volume of data exceeds our ability to decipher its meaning.

Privacy Concerns and Knowledge Discovery

Although not unique to Knowledge Discovery, sensitive information is being collected and stored in these huge data warehouses. Concerns have been raised about what information should be protected from KDD-type access. The ethical and moral issues of invasion of privacy are intrinsically connected to pattern recognition. Safeguards are being discussed to prevent misuses of the technology.

Summary

Knowledge Discovery in Databases is answering a need to make use of the mountains of data that is accumulating daily. KDD enlists the power of computers to assist in the recognizing patterns in data, a task that is exceeds human ability as the size of data warehouses increase. New methods of analysis and pattern extraction are being developed and adapted to KDD. Which method is used depends on the domain and results expected. The accuracy of the recorded data must not be overlooked during the KDD process. Domain specific knowledge assists with the subjective analysis of KDD results. Much attention has been given to the data mining phase of KDD but earlier steps, such as data cleaning, play a significant role in the validity of the results.

The potential benefits of discovery driven data mining techniques in extracting valuable information from large complex databases are unlimited. Successful applications are surfacing in industries and areas were data retrieval is outpacing man's ability to effectively analyze its content. Users must be aware of the potential moral conflicts to using sensitive information.

REFERENCES

(1) Way, J.; and Smith, E.A. "The evolution of Synthetic Radar Systems and Their Progression to the EOS SAR." IEEE Trans. Geoscience and Remote Sensing. Vol 29. No. 6. 1991. Pp962-985.

(2) Fayyad, U.; Simoudis, E.; "Knowledge Discovery and Data Mining Tutorial MA1" from Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95) July 27, 1995 www-aig.jpl.nasa.gov/public/kdd95/tutorials/IJCAI95-tutorial.html

(3) Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; "From Data Mining to Knowledge Discovery: An overview" in Advances in Knowledge discovery and Data Mining. Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; Uthurusamy, R. MIT Press. Cambridge, Mass.. 1996 pp. 1-36

(4) Fayyad, U. "Data Mining and Knowledge Discovery: Making Sense Out of Data" in IEEE Expert October 1996 pp. 20-25

(5) Simoudis, E. "Reality Check for Data Mining" in IEEE Expert October 1996 pp. 26-33

(6) Hand, D. J. 1981 Discrimination and Classification. Chichester, U.K.: John Wiley and Sons

(7) Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; "From Data Mining to Knowledge Discovery: An overview" in Advances in Knowledge discovery and Data Mining. Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; Uthurusamy, R. MIT Press. Cambridge, Mass.. 1996 pp. 1-36

(8) Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P.; "The KKD Process for Extracting Useful Knowledge from Volumes of Data" in Communications of the ACM, November 1996/Vol 39, No.11 pp.27-34

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(9) Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; "From Data Mining to Knowledge Discovery: An overview" in Advances in Knowledge discovery and Data Mining. Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; Uthurusamy, R. MIT Press. Cambridge, Mass.. 1996 pp. 1-36

(10) Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; "From Data Mining to Knowledge Discovery: An overview" in Advances in Knowledge discovery and Data Mining. Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; Uthurusamy, R. MIT Press. Cambridge, Mass.. 1996 pp. 1-36

(11) Simoudis, E. "Reality Check for Data Mining" in IEEE Expert October 1996 pp. 26-33

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Knowledge Management (KM): An Overview

Omankwu, Obinnaya Chinecherem; Nwagu, Chikezie Kenneth, and Inyiama, Hycient

[1] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[2] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

While knowledge management (KM) is becoming an established discipline with many applications and techniques, its adoption in health care has been challenging. Though, the health care sector relies heavily on knowledge and evidence based medicine is expected to be implemented in daily health care activities; besides, delivery of care replies on cooperation of several partners that need to exchange their knowledge in order to provide quality of care. This publication will give an overview of KM, its methods and techniques.

Keywords
Knowledge Management, Data Mining, Knowledge Mining

1. INTRODUCTION

In service base companies, knowledge is a central intangible asset; knowledge management deals with the creation, use, reuse, dissemination of Knowledge. Knowledge Management (KM) became a discipline during the 80's, and the growing role of information technologies enabled the development of efficient KM tools using databases and collaborative software.

KNOWLEDGE MANAGEMENT A

Brief History

Knowledge management had always been a central question in human societies. Indeed, its roots are to be found in the early history of human societies. Philosophers, Western as well as Eastern, have focused their attention on the question of knowledge; already in ancient Greece, 'scientific' discussions often lead to philosophical debates, especially on the concept of knowledge. The creation of epistemology has finally formalized the question of knowledge; indeed, epistemology addresses primarily the question of "what is knowledge?" and discusses its creation and adoption. In the current discipline of knowledge management, philosophical considerations from several schools are taken into account, especially in the ontological knowledge management field (Grenon, 2003).

On the other hand, practical knowledge management has always taken place in the society, and transmission of knowledge was much related to the technical progress. Beginning in the middle age, knowledge transmission occurred under what was called "Wandergesellen" in Germany and "Compagnonnage" in France, where craftsmen and artisan take a tour of the country for 6 months or one year to learn from several masters. This was one of the first structured methodologies for tacit knowledge transmission. Knowledge first spread orally, then in writing; but it was restricted to a low circle of educated people till the development of printing. If the first printing focused on religious and literature purpose, technical and specialized books began to spread after the wide adoption of the printing press.

In the 20th century, management as well as cognitive sciences and psychology led to today's Knowledge Management (KM) (Wiig, 1999). The current situation of KM started in the 1980s with the wide use of information technologies in companies; the focus was on the intangible asset that knowledge represents. The word KM itself appeared in the 80s and the academic discipline was created in 1995 (Stankosky, 2005).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Goals and challenges of KM are many; for companies, KM should increase their performance, help to develop partnerships, evaluate risks, organize management and enhance their economic value. Development of corporate memory and measurement tools also aims at assessing intangible assets in the companies. Besides, knowledge transfer enhancement and companies' performance evaluations became is-sues of major importance.

After twenty productive years in KM, the first criticisms appeared in 2002. T.D. Wilson (Wilson, 2002) discussed the foundation of KM, mainly because of the difficulty to distinguish information from knowledge in most KM theories. He drew the conclusion that KM was a management fad and should disappear in the upcoming years. Actually KM survived well those criticisms, even if the precision of the vocabulary is not comparable to the one used in epistemology or in computer science based KM; probably the reason lies in the real need for companies as well as public organizations to use KM methods.

We can distinguish 2 main KM trends: people and information management centered and in-formation technology centered. We should also recognize two other main orientations, the first is the need of evaluation in terms of performance measurement, and the second is the measurement of knowledge assets in order to evaluate the value of an organization (Wiig, 1999).

KM Frameworks

Frameworks for KM support are based on considerations related to the structure of knowledge and to the structure of organizations where the frameworks are applied. In most of models, knowledge types are determined based on different criteria, such as having structured or unstructured knowledge, and having tacit or explicit knowledge.

First we have to make a distinction between high level frameworks and implementation oriented ones. The latter one focus on the "how to" implement KM in an organization, whereas the first one discuss the question of "what is KM" (Wong & Aspinwall, 2004). As our purpose is to focus more on the "how to" question, we will focus in the next paragraphs on the implementation oriented frameworks.

High level frameworks discuss how to fill the gap between theory and practice,that is the case of Knowledge Creation Frameworks for example (Siebert, 2005).

Nonaka and Takeuchi (Nonaka & Takeuchi, 1995) depict steps to create knowledge in practice that go from perception to representation and from tacit knowledge to explicit one; they also show how those steps can enhance company's efficiency.

Concerning implementation frameworks, Sunassee and Sewry (Sunassee & Sewry, 2002) defined three categories of frameworks: prescriptive, descriptive and hybrids. *Prescriptive* frame-works give direction concerning the procedures that should be used, without describing precisely their content or implementations, for example step approach frameworks are mainly prescriptive frameworks (Wong & Aspinwall, 2004). *Descriptive* frameworks describe the key factors of KM that can drive KM initiatives to success or to failure, their forms of representations are mostly graphical (Wong & Aspinwall, 2004); examples of descriptive frameworks can be found in (Gore

1. Gore, 1999; Holsapple & Joshi, 2002). Finally Hybrid approaches combine both prescriptive and descriptive methods.

It is important to find a way to compare KM frameworks; though, frameworks are dedicated to specific applications which make their comparison complicated. Wong and Aspinwall (Wong & Aspinwall, 2004) proposed a comparison method of frameworks based on four elements, their structure, the knowledge types they represent, the KM processes and the KM influences or factors.

Methods and Techniques in KM

We can categorize the methods and techniques in KM in three groups: people and technology, requirements elicitation and value measurement.

People and Technology

Early approaches of KM frameworks in the early 1990s mainly focused on the structural organization and IT solutions to improve knowledge management (Wiig, 1999). Those methods were adapted for slow moving businesses were goals and technical solutions are perfectly identified and the market does not evolve quickly. But, these approaches were not adapted in a subsequent fast moving business environment where new challenges started arise as fast as they disappear.

Human centered KM has been early identified and became a new school of thought, in the early 1990s. Peters (Peters, 1994) wrote "the answer turns out to lie more with psychology and marketing of knowledge within the family than with bits and bytes". Nowadays frameworks take both *human* and *technical* perspectives into ac-count. We will discuss both approaches separately and show how both are integrated in nowadays frameworks.

Human Perspective: Motivation and Adoption

The main issue for companies is to motivate employees to use KM systems. Not only that the technology matters, but people implication in KM initiatives is a key factor for its success. Without incentives, employees are not ready to share their knowledge; therefore, numerous solutions have been proposed to motivate employees to make use of KM systems. Some companies provide financial incentives (monetary rewards) or non-financial incentives (air miles, days off) for the first users of the KM system. Incentives, financial or not, are particularly efficient in organization where detaining knowledge is often considered as a source of power. In addition to individual incentives, Zand (Zand, 1997) suggests a collaborative win-win reward system, in which a gain for an individual can be a gain for his peers, in opposition with classical win-lose rewards system.

It has also been recognized that higher man-agement should use the system too; Liebowitz (Liebowitz, 1999) cites the success of the KM network of Buckman Labs, which was mostly a result of the high level implication of the senior management and especially the CEO.

The second motivation related issue is *knowledge adoption*; it has been a challenge that people were not ready to use or apply knowledge developed by others. Sussman and Siegal (Sussman & Siegal, 2003) built a theoretical model to understand the underlying issues of knowledge adoption; their study discussed the role of informational influence in the process of knowledge adoption, and showed the importance of the source *credibility* to convince people of the usefulness of the acquired knowledge. Once again, the commitment of senior management, who are trusted in their organization, can have a huge influence on the success of a project.

Technical Perspective: Data Mining, Inference Engine and Multi-Agent Systems

KM tools deals with explicit knowledge, meaning that Knowledge can be written on a support that is mainly an electronic one. Historically, collaborative tools, such as Lotus Notes, have been developed in the 1990s to enhance KM. Recent corporate tools widely adopted Web 2.0 technologies such as wiki platforms, semantic widgets, tagging and so on. Several concepts from the broad computer science research, such as *data mining*, *rules based reasoning*, and *multi-agent systems* have been integrated in KM solutions, the integration of those tools depends on the processes in action.

For instance, computer assisted Knowledge Discovery is mainly based on data mining techniques. A brief look on the papers of the Knowledge Discovery and Data mining (KDD) conference (Li, Liu, & Sarawagi, 2008) - the major conference on Knowledge Discovery - gives an overview of the overwhelming presence of data mining within Knowledge Discovery.

On the other hand, knowledge representation uses ontological models; due to the development of powerful inferences engines. Those representations can be used to infer knowledge from existing one, and shore up Knowledge Discovery processes. Several KM frameworks are based on ontologies (Fensel, 2002; Stojanovic, 2003; Sure, 2002; S.-Y. Yang, Lin, Lin, Cheng, Soo, 2005), since high level representation of Knowledge using ontologies enables powerful queries and Knowledge manipulation, retrieval and discovery.

Finally, the multi-agents system (MAS) paradigm is rightly suited to model the distribution of knowledge on autonomous entities, thus, it is used in order to disseminate knowledge among employees in organizations. MASs also take in account reactivity (adaptation to changes in an environment) and proactivity (anticipation of user needs and consequently taking initiatives). These two factors are the keys for the success of a KM project; indeed, KM initiatives require adapting quickly and being able to handle user needs. In this context, Virtual Knowledge Communities (Maret, Subercaze, & Calmet, 2008) present an efficient way to model KM in organization since it integrates the MAS approach and the ontological representation of Knowledge. Virtual model Knowledge Communities' model has been used for business (Subercaze, Pawar, Maret, & Calmet, 2008) as well as for health care purposes (El Morr, Subercaze, Maret, & Rioux, 2008).

Requirements Elicitation

Requirements can be seen under two angles, a technological one and a human centered one.

From the technological stand point, *storage* of Electronic Knowledge Repository represented a challenge at the early stages of KM; indeed hardware investment can require significant amount of money for a huge amount of data to process. Knowledge Discovery processes also require high computational power; nevertheless with the reduction of hardware costs, storage is no more a critical issue, but the latest research using ontological representation, inference engines, and data mining techniques still required powerful *computational power* (Guo, Pan, & Heflin, 2005).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Value Measurement

Assessing the value of KM is a primary concern for organizations. Like other intangibles assets, the reliability of Knowledge Management measurement in an organization is subject to debate. As underlined in a study for the European Union (Zambon, 2003), internal evaluations based on information provided by managers may be subject to bias and tend to overestimate the value of KM. On the other hand, evaluations conducted by third parties may be imprecise, as third parties may not have access to the internal knowledge assets. The absence of a market for intangible asset can also be a root of evaluation bias; indeed, knowledge as an intangible asset will be evaluated and appear on the financial report but cannot be sold and has no proper market value. Therefore, there is no market structure that can regulate knowledge evaluation.

Several methods have been developed to estimate the value of knowledge in an organization, Skandia is the first company to have dealt with the *Intellectual Capital* (IC) measurement (N. Bontis, 1996). It defined Intellectual Capital as the sum of the human and structural Capital. Human capital combines abilities, knowledge, and innovation potential of the company's employees; it includes the company's philosophy and culture too. This kind of capital is not property of the company, but the company drives benefits out of it. Structural capital is the patents, trademarks, hardware and "everything that gets left behind when employees go home" (Nick Bontis, 2001). IC reports developed by Skandia used 36 metrics to give a monetary value to an organization; metrics includes customer satisfaction, satisfied employees, number of patents, annual turnover. Second generation methods such as *IC-index*, was an extension of the Skandia IC metric, it tried to merge the different indicators of Skandia into a single index (Roos, Roos, Edvinsson, & Dragonetti, 1997). Other metrics were developed to evaluate Knowledge Management Systems (KMS), Kankanhalli et al. (Kankanhalli & Tan, 2004) present a thorough review of KMS metrics.

REFERENCES

Abidi, S. S. R. (2001). Knowledge management in healthcare: towards 'knowledge-driven' decision-support services. *International Journal of Medical Informatics*, *63*(1-2), 5–18.

Ahmad, R., Kausar, A. R., & David, P. (Writer) (2007). *The social management of embodied knowledge in a knowledge community*.

Anderson, R. A., & McDaniel, R. R. (2000). Managing health care organizations: where profes-sionalism meets complexity science. *Health Care Management Review*, *25*(1), 83–92.

Andreas, R. (2005). Three-dozen knowledge-sharing barriers managers must consider. *Jour-nal of Knowledge Management*, *9*(3), 18–35.

Andreas, R., & Nicholas, L. (2006). Knowledge management in the public sector: stakeholder partnerships in the public policy development. *Journal of Knowledge Management*, *10*(3), 24.

Ansell, C. (2007). Fostering Innovation and Col-laboration. *Medical Device Technology*, *18*(1), 52.

Bali, R. K., & Dwivedi, A. N. (Eds.). (2007). *Healthcare Knowledge Management*: Springer.

Batalden, P., & Splaine, M. (2002). What will it take to lead the continual improvement and in-novation of health care in the twenty-first century? *Quality Management in Health Care*, *11*(1), 45.

Bate, S. P., & Robert, G. (2002). Knowledge management and communities of practice in the private sector: lessons for modernizing the National Health Service in England and Wales. *Public Administration*, *80*(4), 643–663.

Bates, D. W., Spell, N., Cullen, D. J., Burdick, E., Laird, N., & Petersen, L. A. (1997). The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *Journal of the American Medical Association*, *277*(4), 307–311. doi:10.1001/jama.277.4.307

Bontis, N. (1996). There's a price on your head: Managing Intellectual Capital Strategically. *Business Quarterly*, (Summer): 40–47.

Bontis, N. (2001). Assessing knowledge assets: a review of models used to measure intellectual capital. *International Journal of Management Reviews*, *3*(1), 41–60.

Brian, D. P. (2006). *Final Report of the Federal Advisor on Wait Times*. Ottawa, Canada: Health Canada.

Buchan, I. E., & Hanka, R. (1997). Exchanging clinical knowledge via Internet. *International Journal of Medical Informatics, 47*(1-2), 39–41.

Burnett, S. M., Williams, D. A., & Webster, L. (2005). Knowledge support for interdisciplin-ary models of healthcare delivery: a study of knowledge needs and roles in managed clinical networks. *Health Informatics Journal*, *11*(2), 146–160.

Caldwell, D. F., Chatman, J., O'Reilly, C. A., III, Ormiston, M., & Lapiz, M. (2008). Implementing strategic change in a health care system: the impor-tance of leadership and change readiness. *Health care management review*, *33*(2), 124(110).

Canadian Health Services Research Foundation. (2003). *The theory and practice of knowledge brokering in Canada's health system*. Ottawa, Canada: Canadian Health Services Research Foundation.

Canadian Institute for Health Information. (2007). *Health Care in Canada*. Ottawa, Ontario: Canadian Institute for Health Information.

Canongia, C., Antunes, A., de Nazare, M., & Pereira, F. (2004). Technological foresight - the

cinc.in (2008). Retrieved October 10, 2008, from http://cyn.in/

Cullen, D. J., Sweitzer, B. J., Bates, D. W., Bur-dick, E., Edmondson, A., & Leape, L. L. (1997). Preventable adverse drug events in hospitalized patients: A comparative study of intensive care and general care units. *Critical Care Medicine*, *25*(8), 1289–1297.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

PERSPECTIVES FOR THE USE OF KM IN HEALTH CARE.

Omankwu, Obinnaya Chinecherem; Nwagu, Chikezie Kenneth, and Inyiama, Hycient

[1] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[2] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

Knowledge management (KM) is becoming an established discipline with many applications and techniques, its perspective view in health care has been challenging. Though, the health care sector relies heavily on knowledge and evidence based medicine is expected to be implemented in daily health care activities; besides, delivery of care replies on cooperation of several partners that need to exchange their knowledge in order to provide quality of care. This publication will give a perspective view of KM in Health care.

Keywords
Knowledge Management, Data Mining, Knowledge Mining

1. INTRODUCTION

In service base companies, knowledge is a central intangible asset; knowledge management deals with the creation, use, reuse, dissemination of Knowledge. Knowledge Management (KM) became a discipline during the 80's, and the growing role of information technologies enabled the development of efficient KM tools using databases and collaborative software.

Beside the current knowledge management roles in the health care sector, few perspectives present an opportunity to develop new health care KM applications. These perspectives are virtual communities, mobility, Electronic Health Record (E.H.R.), and public health.

Virtual Communities

"Virtual" health care providers of different disciplines (e.g. medicine, nursing, social work, physical therapy, etc.) can create teams in which they combine their knowledge and expertise to provide a comprehensive plan of care. Though, it is essential to include patients in virtual health care teams; indeed, patients must be well informed about their conditions, treatment options and how to access them and be actively involved in their treatment (Davis, Wagner, & Groves, 2000). Health Virtual Communities, that include care givers and patients, in order to create manage and coordinate virtual medical teams (Pitsillides, et al., 2004).

Once a Health VC is in place, new knowledge emerges through social interactions (Ahmad, Kausar, & David, 2007). Patients have tacit knowledge about their medical condition and the way they experience their conditions, this tacit knowledge constitute a mine of information for clinical practice; indeed, it allows to get insight into the patient experience and hence assess her/ his quality of life as well as the impact of a drug on a person's life. Virtual communities in this respect constitute an opportunity for a holistic approach to clinical practice.

Besides, Health VCs constitute an opportunity for e-continuing education. In health care, continuous education is essential; some professionals cannot continue practising unless they undergo a yearly continuous education course in order to update their knowledge. In this context, knowledge based Health VCs can play a major role by pro-viding a platform for e-education and knowledge exchange between peers. The creation of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

virtual network of experts opens the road to test new kind of cooperation paradigms and *peer-to-peer e-educational* paradigms (van Dijk, Hugenholtz, A-Tjak, & Schreinemakers, 2006).

Mobility

While managing knowledge will become an important daily practice, the future will be more mobile. We're witnessing already the explosion of mobile interactive devices, mobile health facili-ties, and the proliferation of e-homecare solutions (Hubert, 2006). Mobile knowledge management is the next step in mobile health care situations where patient is away from the point of care (O'Sullivan, McLoughlin, Bertolotto, & Wilson, 2007). The mobility approach is extremely valid in virtual communities (El Morr, 2007; Christo El Morr & Jalal Kawash, 2007; C. El Morr & J. Kawash, 2007); consequently, the creation of mobile Health VCs where knowledge is generated, disseminated and shared by both patients and caregivers is a next step that can provide advantage for both patients and caregivers (Hubert, 2006; Moreno & Isern, 2002; Siau & Shen, 2006).

Electronic Health Record (E.H.R.)

Worldwide, governments are striving to build national wide E.H.R. systems. There has been progress in this direction mainly in developed countries. Once health records are computerized the need will be to reach the right information about a patient at the right time, and to use the E.H.R. data for diagnosis purposes, for personal health decision support, for public health decision support, and for research purposes as well. Though, much of what has been done till now in E.H.R. involves data processing mainly (Van Vlymen, De Lusignan, Hague, Chan, & Dzregah, 2005); besides, health service managers are facing many difficulties when trying to access relevant data routinely for quality improvement (De Lusignan, Wells, Shaw, Rowlands, & Crilly, 2005). KM techniques can play here two roles one for managers and one for practitioners; indeed, KM techniques can help in searching for knowledge (2) in the mass of data gathered helping practitioners to find more effective ways to treat patients by searching for similar patient case histories (O'Sullivan, et al., 2007), and helping managers to get relevant knowledge for total quality management (TQM) (McAdam Leonard, 2001). Establishing, electronic health records, per se, constitute only a first step; using the mass of data gathered in order to support practitioners in generating knowledge and providing quality practise is the challenge ahead.

Evidence-Based Public Health

Networks for health care surveillance continue to evolve (Health Canada, 1999); nevertheless, studies show that information and communication technology are less used in public health than in other sectors of the society (Goddard, et al., 2004; Revere, et al., 2007). Public health is traditionally data processing and data analysis oriented, though there is more awareness that a shift is needed in public health from data driven decision making to knowledge driven decision making, or to put it in Goddart et al. words "pro-vide direct guidance on the relative effectiveness of different interventions in a specific situation" (Goddard, et al., 2004). KM can play a vital role in organizing, structuring and supporting *evidence based public health* decision making (Andreas & Nicholas, 2006; Revere & Fuller, 2008). In this context, research needs to unveil how the public health community communicates and cooperate, particularly in terms of role and communication strategies, artifact used, etc. Different profiles of knowledge health care workers can then be sketched. Research methods from the Computer Supported Collaborative Work (CSCW) field can be used. Findings can well be integrated in the context of Community of Practice where knowledge tools can further knowledge creation, communication and transfer. The medical field is experiencing a move to evidence based medicine, a similar move to *evidence-based public health* is important and would be strategic for an advanced management of population health; KM can play a vital role to make this move.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Knowledge Transfer

Knowledge transfer is concerned with dissemination of knowledge connecting and adapting research findings to the society needs. Increas-ingly, the role of knowledge broker is recog-nized as vital in knowledge transfer (Lind & Persborn, 2000); knowledge brokering "links decision makers with researchers, facilitating their interaction"(Canadian Health Services Research Foundation, 2003; Lomas, 2007). In this context, there is a crucial need to understand how knowledge is transferred, and transformed while it is transferred (Syed, 1999); cognitive theory can be of much help in this domain. This understanding will help providing a feedback to knowledge generators (i.e. researchers) and widen their knowledge (i.e. help generating more knowledge) (Figure 1).

Health 2.0 and Semantic Web

The term health 2.0 embeds the concepts of healthcare, web 2.0 and e-health. Following the web 2.0 principles, health 2.0 is driven by participatory ideas. In health 2.0, each actor of the system, patients, stakeholders are involved in the process of amelioration of the health care system using existing web 2.0 social networking, semantic web and collaborative tools.

As well as web 2.0, health 2.0 is an imprecise term. Most of the applications are focused on enhancing communication in the community. For instance, Sermo1 is physicians community dedicated for information exchange and collaboration', and DoubleCheckMD2 is a patient oriented applications dedicated for drugs side effects; while Vitals3 help patients find a relevant doctor matching search criteria and write reviews on doctors, and PatientsLikeMe4 is an online community for patient with life threatening conditions. In health 2.0, allow patients to share their experience, medical data with other patients, doctors and research organizations; it aims at establishing data-sharing partnerships.

The trend in health 2.0 is to enhance collaborations, either between physicians or between patients, and to create new relationships between patients and doctors and research organizations.

On the other hand, semantic Web technologies enable a next step in accessing data at the scale of the web; indeed, RDF and OWL technologies are being used for knowledge modeling and for large database integrations. Currently, the W3C Semantic Web in Health Care and Life Sciences Interest Group (HCLSIG) aims at offering a better access.

to information from many domains and processes for an efficient decision support and disease management. Initiatives like OBI (Ontology for Biomedical Investigations) or RNA Ontology Consortium are the results of the movement initiated by the HCLSIG.

While, the current health 2.0 applications are based on relational databases; we believe that in the near future, we will see a merger between health 2.0 and the semantic web technologies developed by HCLSIG. The resulting applications could fairly improve automated knowledge management related to healthcare.

CONCLUSION

Knowledge management in health care is progressing; the complexity and challenges facing the health care sector can be addressed by adopting of KM strategies.

The use of KM in health care is promising to enhance the quality of care for patients by providing them with a continuity of care. The implementation of Health care KM system will allow health care partners (e.g. practitioners, administrators, etc.) to conduct evidence based practice and to collaborate relying on the best knowledge available.

This is a challenge that opens the way to more innovations in both KM and health. The current state of KM in health care can be improved; we believe that new practices such as, health 2.0 applications, VCs and evidence based medicine will help to increase the global quality of care of the patients as well as the efficiency of KM in healthcare.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

Abidi, S. S. R. (2001). Knowledge management in healthcare: towards 'knowledge-driven' decision-support services. *International Journal of Medical Informatics*, *63*(1-2), 5–18.

Ahmad, R., Kausar, A. R., & David, P. (Writer) (2007). *The social management of embodied knowledge in a knowledge community*.

Anderson, R. A., & McDaniel, R. R. (2000). Managing health care organizations: where profes-sionalism meets complexity science. *Health Care Management Review*, *25*(1), 83–92.

Andreas, R. (2005). Three-dozen knowledge-sharing barriers managers must consider. *Jour-nal of Knowledge Management*, *9*(3), 18–35.

Andreas, R., & Nicholas, L. (2006). Knowledge management in the public sector: stakeholder partnerships in the public policy development. *Journal of Knowledge Management*, *10*(3), 24.

Ansell, C. (2007). Fostering Innovation and Col-laboration. *Medical Device Technology*, *18*(1), 52.

Bali, R. K., & Dwivedi, A. N. (Eds.). (2007). *Healthcare Knowledge Management*: Springer.

Batalden, P., & Splaine, M. (2002). What will it take to lead the continual improvement and in-novation of health care in the twenty-first century? *Quality Management in Health Care*, *11*(1), 45.

Bate, S. P., & Robert, G. (2002). Knowledge management and communities of practice in the private sector: lessons for modernizing the National Health Service in England and Wales. *Public Administration*, *80*(4), 643–663.

Bates, D. W., Spell, N., Cullen, D. J., Burdick, E., Laird, N., & Petersen, L. A. (1997). The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *Journal of the American Medical Association*, *277*(4), 307–311. doi:10.1001/jama.277.4.307

Bontis, N. (1996). There's a price on your head: Managing Intellectual Capital Strategically. *Business Quarterly*, (Summer): 40–47.

Bontis, N. (2001). Assessing knowledge assets: a review of models used to measure intellectual capital. *International Journal of Management Reviews*, *3*(1), 41–60.

Brian, D. P. (2006). *Final Report of the Federal Advisor on Wait Times*. Ottawa, Canada: Health Canada.

Buchan, I. E., & Hanka, R. (1997). Exchanging clinical knowledge via Internet. *International Journal of Medical Informatics*, *47*(1-2), 39–41.

Burnett, S. M., Williams, D. A., & Webster, L. (2005). Knowledge support for interdisciplin-ary models of healthcare delivery: a study of knowledge needs and roles in managed clinical networks. *Health Informatics Journal*, *11*(2), 146–160.

Caldwell, D. F., Chatman, J., O'Reilly, C. A., III, Ormiston, M., & Lapiz, M. (2008). Implementing strategic change in a health care system: the impor-tance of leadership and change readiness. *Health care management review*, *33*(2), 124(110).

Canadian Health Services Research Foundation. (2003). *The theory and practice of knowledge brokering in Canada's health system*. Ottawa, Canada: Canadian Health Services Research Foundation.

Canadian Institute for Health Information. (2007). *Health Care in Canada*. Ottawa, Ontario: Canadian Institute for Health Information.

Canongia, C., Antunes, A., de Nazare, M., & Pereira, F. (2004). Technological foresight - the

cinc.in (2008). Retrieved October 10, 2008, from http://cyn.in/

Cullen, D. J., Sweitzer, B. J., Bates, D. W., Bur-dick, E., Edmondson, A., & Leape, L. L. (1997). Preventable adverse drug events in hospitalized patients: A comparative study of intensive care and general care units. *Critical Care Medicine*, *25*(8), 1289–1297.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Robotics and Artificial Intelligence: Differences and Similarities

Omankwu, Obinnaya Chinecherem, Nwagu, Chikezie Kenneth, and Inyiama, Hycient

[1] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[2] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

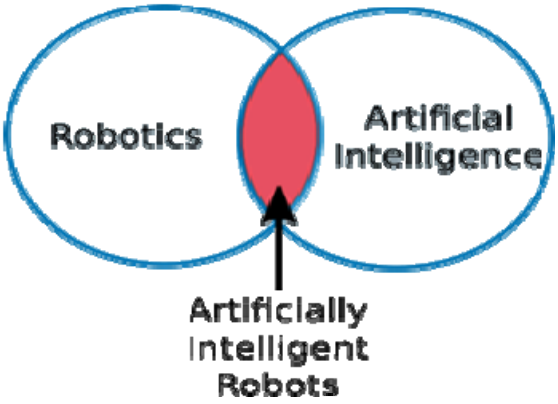
Robotics and artificial intelligence serve very different purposes. However, people often get them mixed up. Robotics is a branch of technology which deals with robots. Robots are programmable machines which are usually able to carry out a series of actions autonomously, or semi-autonomously**.** Artificial intelligence (AI) is a branch of computer science. It involves developing computer programs to complete tasks which would otherwise require human intelligence.

Keywords: Robotics, artificial intelligence, algorithm.
.

Introduction

Robotics and artificial intelligence serve very different purposes. However, people often get them mixed up.

Are Robotics and Artificial Intelligence the Same Thing? The first thing to clarify is that robotics and artificial intelligence are not the same thing at all. In fact, the two fields are almost entirely separate. A Venn diagram of the two would look like this:



I guess that people sometimes confuse the two because of the overlap between them: Artificially Intelligent Robots.

To understand how these three terms relate to each other, let's look at each of them individually.

What Is Robotics?

Robotics is a branch of technology which deals with robots. Robots are programmable machines which are usually able to carry out a series of actions autonomously, or semi-autonomously.

In my opinion, there are three important factors which constitute a robot:

1. Robots interact with the physical world via sensors and actuators.
2. Robots are programmable.
3. Robots are usually autonomous or semi-autonomous.

Robots are "usually" autonomous because some robots aren't. Telerobots, for example, are entirely controlled by a human operator but telerobotics is still classed as a branch of robotics. This is one example where the definition of robotics is not very clear.

It is surprisingly difficult to get experts to agree exactly what constitutes a "robot." Some people say that a robot must be able to "think" and make decisions. However, there is no standard definition of "robot thinking." Requiring a robot to "think" suggests that it has some level of artificial intelligence.

However you choose to define a robot, robotics involves designing, building and programming physical robots. Only a small part of it involves artificial intelligence.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

What Is Artificial Intelligence?

Artificial intelligence (AI) is a branch of computer science. It involves developing computer programs to complete tasks which would otherwise require human intelligence. AI algorithms can tackle learning, perception, problem-solving, language-understanding and/or logical reasoning.

AI is used in many ways within the modern world. For example, AI algorithms are used in Google searches, Amazon's recommendation engine and SatNav route finders. Most AI programs are not used to control robots.

Even when AI is used to control robots, the AI algorithms are only part of the larger robotic system, which also includes sensors, actuators and non-AI programming.

Often — but not always — AI involves some level of machine learning, where an algorithm is "trained" to respond to a particular input in a certain way by using known inputs and outputs. We discuss machine learning in our article Robot Vision vs Computer Vision: What's the Difference?

The key aspect that differentiates AI from more conventional programming is the word "intelligence." Non-AI programs simply carry out a defined sequence of instructions. AI programs mimic some level of human intelligence.

What Are Artificially Intelligent Robots?

Artificially intelligent robots are the bridge between robotics and AI. These are robots which are controlled by AI programs.

Many robots are not artificially intelligent. Up until quite recently, all industrial robots could only be programmed to carry out a repetitive series of movements. As we have discussed, repetitive movements do not require artificial intelligence.

Non-intelligent robots are quite limited in their functionality. AI algorithms are often necessary to allow the robot to perform more complex tasks.

Let's look at some examples.

Example: Non-Artificially Intelligent Cobot

A simple collaborative robot (cobot) is a perfect example of a non-intelligent robot.

For example, you can easily program a cobot to pick up an object and place it elsewhere. The cobot will then continue to pick and place objects in exactly the same way until you turn it off. This is an autonomous function because the robot does not require any human input after it has been programmed. However, the task does not require any intelligence.

Example: Artificially Intelligent Cobot

You could extend the capabilities of the cobot by using AI.

Imagine you wanted to add a camera to your cobot. Robot vision comes under the category of "perception" and usually requires AI algorithms.

For example, say you wanted the cobot to detect the object it was picking up and place it in a different location depending on the type of object. This would involve training a specialized vision program to recognize the different types of object. One way to do this is using an AI algorithm called Template Matching, which we discuss in our article How Template Matching Works in Robot Vision.

Conclusion

As you can see, robotics and artificial intelligence are really two separate things. Robotics involves building robots whereas AI involves programming intelligence. Software robot" is the term given to a type of computer program which autonomously operates to complete a virtual task. They are not physical robots, as they only exist within a computer. The classic example is a search engine webcrawler which roams the internet, scanning websites and categorizing them for search. Some advanced software robots may even include AI algorithms. However, software robots are not part of robotics.

References

[1] J.J. Craig, *Introduction to Robotics*. Reading, MA: Addison-Wesley, 2nd ed., 1989.

[2] L. Zhenyu, C. Yinglin, and Q. Daokui, “Research on robot calibra-tion,” *Robot*, vol. 24, no. 5, pp. 447–450, 2002.

[3] S. Lei, L. Jingtai, S. Weiwei, W. Shuihua, and H. Xingbo, “Geometry-based robot calibration method,” in *Proc. IEEE Int. Conf. Robotics and Automation*, pp. 1907–1912, 2004.

[4] O. Khatib, “Real-time obstacle avoidance for manipulators and mobile robots,” *Int. J. Robot. Res.*, vol. 5, no. 1, pp. 90–98, 1986.

[5] J.F. Canny, *The Complexity of Robot Motion Planning*, Cambridge, MA: MIT Press, 1988.

[6] J.T. Schwartz and M. Sharir, “On the ‘piano movers’ problem: 1. The case of two-dimensional rigid polygonal body moving amidst polygo-nal barriers,” *Commun. Pure Appl. Math.*, vol. 36, pp. 345–398, 1983.

[7] J. Bobrow, S. Dubowsky, and J. Gibson, “Time-optimal control of robotic manipulators along specified paths,” *Int. J. Robot. Rese.*, vol. 4, no. 3, pp. 3–17, 1985.

[8] J.J.E. Slotine and W. Li, *Applied Nonlinear Control*. Upper Sadlle Rikver, NJ: Prentice-Hall, 1991.

[9] K.J. Aström and K.B. Wittenmark, *Adaptive Control*. Reading, MA: Addison-Wesley, 1989.

[10] B. Siciliano and L. Villani, *Robot Force Control*. Norwelll, MA: Kluwer, 1999.

[11] P.C. Watson, “Remote center compliance system,” U.S. Patent No. 4098001, Jul. 1978.

[12] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* 2nd ed. Uppper Saddle River, NJ: Prentice Hall, 2003.

[13] R.S. Michalski, J.C. Carbonell, and T.M. Mitchell, *Machine Learning*, Palo Alto, CA: Tioga, 1983.

[14] S.G. Tzafestas, “Fuzzy systems and fuzzy expert control: An overview,” *Knowledge Eng. Rev.*, vol. 9, no. 3, pp. 229–268, 1994.

[15] G. Niemeyer and J.J.E. Slotine, “Telemanipulation with time delays,” *Int. J. Robot. Res.*, vol. 23, no. 9, pp. 873–890, 2004.

[16] B. Davies, *A Discussion of Safety Issues for Medical Robots*, In R. Taylor, S. Lavallee, G. Burdea, and R. Moesges, Eds. *Computer-Integrated Surgery*. Cambridge, MA: MIT Press, pp. 287–296, 1996.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[17] M. Blackwell, C. Nikou, A. DiGioia, and T. Kanade, “An image over-lay system for medical data visualization,” *Med. Image Anal.*, vol. 4, pp. 67–72, 2000.

[18] R. Kumar, P. Berkelman, Gupta, A. Barnes, P.S. Jensen, L.L. Whitcomb, and R.H. Taylor, “Preliminary experiments in cooperative human/robot force control for robot assisted microsurgical manipula-tion,” in *Proc. IEEE Int. Conf. Robotics Automation*, pp. 610–617, 2000, San Francisco, CA.

[19] G.H. Ballantyne and F. Moll, “The da vinci telerobotic surgical system: The virtual operative field and telepresence surgery,” *Surg. Clin. North Amer.*, vol. 86, no. 6, pp. 1293–1304, 2003.

[20] J. Marescaux and F. Rubino, “The zeus robotic system: Experimental and clinical applications,” *Surg. Clin. North Amer.*, vol. 86, no. 6, pp. 1305–1315, 2003.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Faculty assignment and timetabling using optimisation

MA Nang Laik
Senior Lecturer, School of Business
Singapore University of Social Sciences
Singapore
E-mail: nlma@suss.edu.sg*

SEN Prabir
Director Research, STATGRAF Research
Vancouver, Canada
Email: sen.prabir@gmail.com

***Abstract*— Resource planning in university is a very hard management science problem. Faculty members are expensive resource that a university needs to utilize them efficiently and deploy them effectively for courses that they can teach. In this paper, we focus on one of the most important problems in the universities – the academic calendar which comprised of faculty-course assignment, course scheduling and timetabling. We propose an innovative two-steps approach to solve the problem using mathematical models to optimize the resource allocation while satisfying the faculty preferences. We also showcase using a real-world example how this problem is solved easily and solution improves the productivity of the staff and enhances the satisfaction of faculty.**

Keywords—faculty assignment, timetable, university, minimze cost, class schedule

I. INTRODUCTION

Universities today are very complex and they serve thousands of students annually ranging from undergraduates, post-graduates to executives training programmes. The mission of the university in this study is to provide lifelong education, equipping learners to serve society in a positive manner. For universities, faculty members (or "faculty") are the most expensive resources. Faculty involves in course development, assessment writing, and course delivery and grading. In addition, they also need to support other administrative tasks such as student recruitment, information sharing session, applied project supervision and industrial relationship. Some of them focus on research studies that require collaboration with the industry as an integral part of the faculty duties to keep in touch with the industry, understand the market needs and enhance the quality of classroom teaching. In the meantime, if there is an internship opportunity from the industry, students attend interviews and perform short-term internship in the company, for say ten weeks, or, if long-term, for six months.

There are two semesters in the university, one run from August to December which is referred as "term 1" and January to June as "term 2". Before the start of each semester, the administrative team needs to forecast the demand for each course and determine the required number of sessions for each course. Course is a series of classes, on a particular subject. A faculty may teach one session of the course, or teach multiple session of the same course. University offers some courses only once a year and others in every term. Some of the courses are foundation courses and, thereby the demand for such courses is relatively higher than the elective course, which may only have a single session. We also have a minimum class size of 20 and maximum class size of 50. If the demand falls below 20, university no longer offers the course to avoid sub-optimal use of resources. We also focus on the class interactivity thus we limit the number of students in the class to be greater than or equal to 20 but less than or equal to 50.

Faculty includes full-time employees of the university who hold Ph.D or equivalent doctorate degree. Faculty assignment problems is a non-trivial problem as the faculty are not homogenous. They can only teach those courses that are in their area of experts or fields of study, and not all other courses, as the mismatch skills and knowledge which is not ideal for both faculty and students. Some professionals or instructors from other institutions or industry who are associated with the university (e.g., by teaching some courses or supervising students) but do not hold professorships may be appointed as adjunct faculty or associates. There is a large pool of associates who teach courses in their area of expertise. University deploys them for unfulfilled courses by the full-time faculty. In any case, the associates are part-time staffs on contract and they are not entitled to receive any medical and HR benefits like the full-time faculty.

The faculty-course allocation (i.e. faculty assignment) is an important decision making process for the associate dean of the school with his administrative team need to make sure that all the course demands are met. All the courses are assigned to the faculty members (Full-time and part-time) who can teach the course well. The team also needs to ensure that all the full-time faculty members are fully-utilized before engaging the associates. This is done manually purely based on the historical teaching record and most of the times, faculty preferences are not taken into consideration. This raises a lot of concerns and dissatisfaction among the faculty members. The administration team also need to schedule the course to various timeslots based on the availability of the classroom. This is timetabling problem.

The stated problem, faced by associate dean and administrative team every semester, is done manually now. It is mundane and involve a lot of man-hours to complete the task. Thus, we have proposed a decision support system which can plan the faculty

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

assignment and timetabling automatically with minimal human intervention taking into consideration of operational constraints.

In the next few sections, we discuss the literature review; the important decision making process at one of the private universities in Singapore; data analysis to identify the issue of using the "gut-feel" and experience to do the resource planning at the university. We propose two-steps optimization model for faculty assignment and timetabling to meet the required demand and run the programme. The approach yields better results and improves the faculty satisfaction level as they are happy with the course allocation and schedule. We also present some computational result and performance of the models. Finally, we discuss about limitation of the model proposed and future work to be done as a conclusion.

II. Literature Review

Generating academic calendar for a university is a very challenging and time consuming task due to diverse demands, faculty preference and availability, limited time slots and class rooms' requirement. Although, there are other assignment problems in a university, academic calendar problem is the most frequent and scheduling classes is the most challenging one. The academic calendar problem consists two kinds of assignment problems. One is the faculty assignment: to determine which faculty is assigned to the course based on their expertise and preference. And, the other is course scheduling: to determine the correspondence between class-timeslot and classrooms. The context and complexity of the assignment problems are dependent on the relevant university systems and various approaches have been proposed for the problem and solved by various researchers [1, 2, 3, 4, 5, 6 and 7].

Adewumi, Sawyerr, and Ali [2] addressed lecturer scheduling at a Nigerian University, and uses an iterative process to generate schedules based on the degree of violation of hard constraints. Daskalaki and Birbas, [5] developed a two-stage procedure for a department providing structured curricula for well-defined groups of students. The procedure includes a relaxation approach for computationally heavy constraints, and sub-problems to obtain timetables for each day of the week. Derigs and Jenal [7] described a Genetic Algorithm – based system for professional course scheduling using strategies such as pre-assigning subsets of courses. Dinkel Mote and Venkataramanan [9] used a network-based model considering the dimensions of faculty, subject, time, and room for the College of Business Administration at Texas A&M University. Other articles describing heuristic approaches to course scheduling in university environments include [4, 6, 8 and 10].

Fong, Asmuni, McCollum, McMullan and Omatu [11] proposed a new hybrid method which is a combination of a great deluge and artificial bee colony algorithm (INMGD-ABC) to solve the university timetabling problem. Artificial bee colony algorithm (ABC) is a population based method that has been introduced in recent years and has proven successful in solving various optimization problems effectively. However, as with many search based approaches, there exist weaknesses in the exploration and exploitation abilities which tend to induce slow convergence of the overall search process. Therefore, hybridization is proposed to compensate for the identified weaknesses of the ABC.

Gunawan, Ng and Poh [12] developed a mathematical model to solve teacher assignment and course scheduling for a master course. An initial solution is obtained by a mathematical programming approach is based on Lagrangian relaxation. This solution is further improved by a simulated annealing algorithm. The proposed method has been tested on instances from a university in Indonesia, as well as on several randomly generated datasets, and the corresponding computational results are reported.

Hinkin and Thompson [13] considered integrated teacher assignment and course scheduling at a university in Indonesia, and used a heuristic based on Lagrangian relaxation. The models were solved in phases using CPLEX[8]. The authors developed a computer program to automate the scheduling process, considering conflicts among core required courses, and among electives within areas. The program was used by an administrator in the student services office.

Koide [14] developed a prototype system for the examination proctor assignment in Konan University by reference to the mathematical modeling by [16]. They focused on the proctor assignment and the target model considered some different types of constraints with respect to workload in a day from the constraints in [16] model. A mixed integer programming model was proposed and an optimal solution was derived through CPLEX [8], commercial optimization software. The resulting assignment sounded acceptable for the registrar staffs nevertheless some additional practical conditions were neglected for simplification of the mathematical model. This study extends the previous model and discusses the usefulness of the system for system users in the practical assignment task. The timetabling problem is generally large, highly constrained, and solution by exact optimization methods is difficult [15].

Onouchi, Uchigaito Sasaki M. [16] studied timetabling problem for final examinations in their university. The problem was solved in two stages; examination timetabling and classroom assignment were conducted in the first stage and proctor assignment in the second stage. The problems in both stages were formulated as mixed integer programming (MIP) and solved by using commercial optimization software. The authors proposed meta-heuristic approach as it is more easily comprehensible for system users.

Tim Roughgarden [17] discussed the problem of optimizing the performance of a system for the concrete setting of scheduling "centrally controlled" jobs and formulated this goal as an optimization problem via Stackelberg games, games in which one player acts a leader (here, the centralized authority or academic dean interested in optimizing system performance)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and the rest as followers (the faculty members). The problem is then to compute a strategy for the leader (a Stackelberg strategy) that induces the followers to react in a way that (at least approximately) minimizes the total latency in the system.

Based on the literature review, we have identified that faculty assignment and course-scheduling or timetabling is a very complex yet critically important to the university. Most researchers develop mathematical models and solve it using search algorithms or heuristics to achieve the goal. The contribution of our paper is as followed, we develop a two-steps approach to solve the faculty assignment and timetabling sequentially and finally use the output from both models to get the faculty schedule. This approach reduces the computation time tremendously as the number of variables are greatly reduce using our approach.

III. Problem description

The university discussed here is a new university which has just started operating recently. In each year, for performance appraisal, university ask faculty members to put in their preferences (first, second, third choices) in term of courses to teach. Full-time faculty members teach six classes in an academic year as per the employment contract; however, they may get some reduction in teaching load if they involve in course development or significant administrative work to support the programme. Their performances are measured by substantive contributions to the learning of their students and to their field, as well as to make service contributions to their field and the university. Therefore, student's feedback along with effectiveness of course delivery, quality of course development and effectiveness of mentoring students are essential for their career development and mentoring to ensure accountability and equity across the faculty. In essence, their performance indicators are of two kinds: 1) those that denote scholarly activities relevant to a performance area and 2) those that provide service support of the quantity and quality of content delivery activity in a performance area. Neither the number of activities nor the number of supporting services necessarily indicates a high (or low) quality of performance; instead, consider a combination of quantitative and qualitative elements when evaluating performance.

The adjunct faculty members are part-time staff and they need to fill in a form to indicate the courses that they are eligible to teach. Due to some human resource policy and regulation, adjunct faculty members teach no more than two classes in a year. By the end of term 2, the administrative team has the estimated number of students enrolled for each course before the new academic year begins in August. Based on the student enrollment, they compute the number of sessions required for each course. Each class needs a minimum class size of 20 students and a cap of 50 due to the classroom capacity. If there are 120 students signed up for the course, for example, then the total number of sessions required is 3 and they try to balance the number of students in each class. Faculty are assigned to courses based on their expertise and the current objective is to ensure that all the courses demand are being fulfilled taking into considering of faculty preference. It is a multi-faceted decision making process for which a system will be required to solve it effectively to assign faculty to courses that they are able to teach and utilize them effectively.

After the faculty assignment, the administrative team works on scheduling the classes. There are only fifteen time-slots available in a week. And, there are three time-slots daily, we referred as "t1 or morning" for 08:30am to 11:30am, "t2 or afternoon" for 12:00pm to 3:00pm and lastly "t3 or evening" for 3:30pm to 6:30pm. Since there are a limited number of classrooms available in the university, classes are assigned to a timeslot, based on classroom availability. Therefore, one of the objectives is also to schedule all the courses to a timeslot and to an available classroom. At the end of this process, the timetable for all the course which denotes day of week, timeslot and the classroom, will be available.

Once courses allocated and classes scheduled, the faculty members who teach the course are assigned to the schedule. If a member of a faculty has conflict with the allocated schedule, the member is allowed to internally swap with another faculty member within the course. A faculty member cannot teach more than one course at the same time-slot of the day. The final output is the faculty teaching schedule.

The university would like to harness on the power of analytics, to develop a decision support system which automatically allocate, assign and schedule in that sequential order. The proposed solutions aide them in their decision making and achieve an optimal outcome.

IV. Modeling & Compuational result

We model the faculty assignment problem as an integer programming (IP) model. The objective for the model is to minimize the cost. The cost of allocating a course to a faculty is partially based on the preferences list. Each faculty member is asked to submit three preference courses each year. The model aims to assign the course to the faculty's preference as far as possible so as to increase the satisfaction level of the staff. We assign lower cost if the course is in the preference list. The lowest cost is assigned to the first preference, which is followed by second and third choices. Example: we may assign 100 to first choice, followed by 200 to second choice and 300 to third choice. Higher penalty is applied for assigning a course to a faculty member outside of the preference list. It is also more expensive to engage an adjunct than the full-time faculty, therefore treat them separately from now on.

Let i be the number of courses, i = 1, 2, 3, ... , n
Let j be the number of faculty, j = 1, 2, 3, …, m
Let x_{ij} be the decision variable, where course i is allocated to faculty j
Let d_i be the demand for the course i

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Let f_j be the number of sessions required to teach for faculty j
Let c_{ij}be the cost of assigning course i to faculty j

Problem P1 – Faculty assignment

Objective: $\min \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij}$ (1)
$\sum_{j=1}^{m} x_{ij} \geq d_i, \forall i = 1, 2, 3, \dots, n$ (2)
$\sum_{i=1}^{n} x_{ij} \leq f_j, \forall j = 1, 2, 3, \dots, m$ (3)
$x_{ij} \in Z^+$ (4)

The first equation is the objective function, we want to minimize the cost of assignment which is translated to meeting the preferences of each faculty as much as possible while meeting all the operational constraints.

The equation (2) stated that number of faculty assigned for each course must be larger than or equal to the demand for the course. Since it is a minimization problem, the system assigns the minimum classes to the faculty. Equation (3) noted that the number of classes assigned to each faculty i, must be less than the required number of teaching sessions required. For normal full-time faculty, $f_j$is 6 and for adjunct faculty, f_j is 2.

This mathematically model is compact. The challenge that we face during the modeling is to assign appropriate cost to c_{ij}. We have taken into consideration of faculty's preference, faculty's expertise as well as the cost of conducting a course by full-time and adjunct. The cost of conducting a course for full-time faculty also varies according to the rank. Currently there are three ranks in the university, they are namely assistant professor, associate professor and professor ranks. The teaching cost for an assistant professor is the lowest, followed by 30% increased for associated professor and 50% increased for professor. These are users input and the users can change these inputs according to university pay-structure.

After this assignment, the team initiates timetabling process by scheduling the courses to timeslots where there are available classrooms. There are some basic assumptions for this model. The class schedule/timetable for the course is the same weekly throughout the semester, which mean that if a course BA100 is scheduled on Monday 8:30am and the venue is room 4015, then the same schedule is valid for each week within a semester. The courses can be scheduled for three timeslots: (t1, t2 t3) from Monday to Friday, total of 15 timeslots, T, in a week. In this case, we try to find the best time slot to schedule classes each week according to classrooms availability. The schedule for a class remains the same for the whole semester, unless they fall on a public holiday and, in that case, we find a make-up class on Saturday of the same week. Since the university is expanding, the classroom facilities become scarce resource. We schedule a course at a particular timeslot and to an available classroom. Thus, the objective is to schedule all the courses to timeslots and available classrooms. After this schedule process, then we assign faculty members who teach the course to a timetable. If a faculty member has a conflict with the scheduled timetable, then we allow them to internally swap with another faculty member from the same course. A faculty member cannot teach more than one course at the same timeslot of the day, they should be allocated different timeslots. Timetable is a collection of timeslots for a week, for a course or for a faculty member.

Let i be the number of courses, i = 1, 2, 3, ... , n
Let t be the number of timeslots available, t = 1, 2, 3, …,T
Let y_{it} be the decision variable, where course i is allocated to time t
Let d_i be the demand for the course i
Let r_t be the number of classrooms available
Let c_{it}be the cost of assigning course i to timeslot t

Problem P2 – Timetabling

Objective: $\min \sum_{i=1}^{n} \sum_{t=1}^{T} c_{it} y_{it}$ (5)
$\sum_{t=1}^{T} y_{it} \geq d_i, \forall i = 1, 2, 3, \dots, n$ (6)
$\sum_{i=1}^{n} y_{it} \leq r_t, \forall t = 1, 2, 3, \dots, T$ (7)
$y_{it} \in \{0,1\}$ (8)

The equation (5) is the objective function, we want to minimize the cost of assigning course i to timeslot t, if we want to schedule more afternoon classes provided there are classroom available, then we can assign less cost to noon timeslot each day. It is generally true that students prefer noon classes than early morning class or evening class.

The equation (6) stated that number of timeslots allocated to each course must be larger than or equal to the number of classes required for the course. Equation (7) noted that the number of classes assigned in each time slot t, must be less than the classroom available. Finally, the decision variable, y_{it} is a binary variable; it is assigned 1 if course i is assigned to time t or 0 otherwise. Assuming, for C1 courses there are 3 classes weekly. The weekly class schedule can be on t1 (Monday morning), t11 (Thur noon), t14 (Fri noon).

Finally, we match the faculty assignment to the course timetable using some basic rules and we get the faculty schedule. We ensure that one faculty can't teach two classes at the same timeslot. Otherwise, we perform some minor adjustments.

The model can be solved very quickly using Excel solver if the problem is not big or using some commercial software like IBM CPLEX Optimizer. We have used past university data and run the model for 1 year. For the 20 courses, 80 classes and 16 faculty, we are able to solve the problem in less than 30 minutes. It greatly improve the productivity of the administrative team and they just need to look at the preliminary result and made some minor adjustments when required.

We use the following example to illustrate how our approach works. Assuming we have 10 courses, 40 classes and 5 full-time faculty and 5 part-time adjunct faculty. The faculty has given their preferences as well as courses that they can teach in advance. There are 15 timeslots in a week namely, t1, t2 .. t15.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

t1 represents Monday morning, t2 represents Monday noon and t15 for Friday evening. There are also only three classrooms available to book for each timeslot. What is the faculty assignment, timetable for the courses and faculty timetable? Using the above formulation and using Excel solver, we are able to solve the problem in minutes. Here are some sample output from the model.

This formulation has some limitations. The model may allocate different courses to faculty, e.g. they need to teach four different courses in a year which is neither desirable as one faculty shouldn't teach more than three different courses in a year, nor optimal for the program. But the occurrence of this allocation is very rare as the model try to assign the course within the faculty preference list of three choices.

Assuming that we have 10 courses, 10 faculty member to assign in 15 timeslot, the number of variables to solve all of them in one model is 10x10x15 = 1500 variable. If we split the problem into two-steps approach, the total number of variables for problem P1 is 10x10 = 100 and P2 is 10x15 =150, which is much lower than 1500. If we increases the number of courses to 100 courses, 100 faculty and 15 timeslot, the number of variables become very large and computation time increases exponentially.

Next, we are going to use a simple example to illustrate how our model work in real-world.

For a small business programme in our university, we only have 5 full-time faculty members and 5 adjuncts. We are going to offer 10 courses in the coming year and based on the past demand, we know that the number of classes we need to be offered is 40. Refer to the input Table 1 below for detail. For example, C10 is the foundation course, we need to offer 10 sessions of C10 in a year, but C3 is an elective so there is only 1 session offered. Each faculty is asked to fill in a table to state their preference as well as up to 3 other differences courses that they can teach based on their expertise. Adjunct faculty are asked to also fill in up to 3 courses that they can teach. These form the input (Table 1) for our mathematically model.

Table 1: Input for example:

Course demand:

course	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
demand	3	2	1	2	6	3	4	5	6	8

Faculty preferences and courses they can teach:

Resources	1st	2nd	3rd	Courses you can teach		
R1	C5	C3	C2	C1	C4	C7
R2	C1	C4	C5	C2	C6	C9
R3	C3	C8	C6	C4		
R4	C6	C8	C7	C10	C2	
R5	C2	C1	C5	C6	C8	

Resources	Courses that adjuncts can teach		
R6	C1	C3	C5
R7	C2	C4	C8
R8	C1	C5	C8
R9	C3	C6	C9
R10	C5	C7	

Number of classrooms available: 3

Timeslots: t1, t2, t3, … , t15

Table 2: Output from P1

Resources	Courses assigned
R1	C2(1), C5(1), C7(4)
R2	C1(2), C9(4)
R3	C4(2), C6(1), C8(3)
R4	C6(2), C8(2), C10(2)
R5	C2(1), C5(5)
R6	C3(1), C10(1)
R7	C10(2)
R8	C1(1), C10(1)
R9	C9(2)
R10	C10(2)

We can derive the courses that each faculty needs to teach from the table above and share with the faculty who are teaching the courses – refer to Table . We can verify that we assign all the courses and sessions to faculty refer to their preference list. The faculty are also aware that the system will try its best to find the best match but it may not be able to fulfill all their wishes as the most important problem is to satisfy the operational constraints to ensure that all courses are assigned at the end of this stage with minimum cost.

After we have run **P1** model, we continue with the class timetabling problem **P2**. For timetabling, we need to assign a class room and timeslot for each class subjected to room availability. The output of this model P2 is a weekly timetable for all the courses and their respective classroom. A faculty teaches a course, which consists of multiple session of weekly classes over the whole semester. A class timetable denotes day of week and timeslot in the designated class room weekly, thus we will use class and course interchangeable for this purpose.

In this example, there are only 3 class room available at each time slot and there are altogether 15 time slots available each week. The output from P2 is shown in

Table 2.

Using the output from **P1** and **P2**, we need to schedule the faculty timetable. This is done using a simple rule. One faculty can only teach one class at any point in time, if there is a clash, we swap the time table with another class.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
t1	0	0	0	0	0	0	0	0	1	1
t2	0	0	0	0	1	0	0	0	1	1
t3	0	0	0	0	0	0	0	1	1	1
t4	0	0	0	0	0	0	0	0	1	1
t5	0	0	0	0	1	0	1	0	0	1
t6	0	0	0	0	0	0	0	1	1	1
t7	0	0	0	0	0	0	0	0	1	1
t8	0	0	0	0	0	1	0	1	0	1
t9	0	1	0	0	1	0	0	1	0	0
t10	0	0	0	0	0	0	1	1	0	0
t11	0	0	0	1	1	1	0	0	0	0
t12	0	0	0	0	1	1	1	0	0	0
t13	1	0	0	1	0	0	1	0	0	0
t14	1	0	0	0	1	0	0	0	0	0
t15	1	1	1	0	0	0	0	0	0	0

Table 2: Output from P2 - (Timetabling)

Table 3: Matching the faculty assignment with schedule:

R1	C2 (1) t9	C5 (1) t2	C7 (4) t5,t10,t12,t13
R2	C1 (2) t13, t14	C9 (4) t1, t2, t3, t4	
R3	C4 (2) t11, t13	C6 (1) t8	C8 (3) t3, t6, t9 (t8 clash so change to t9)
R4	C6 (2) t11, t12	C8 (2) t8, t10	C10 (2) t1, t2
R5	C2 (1) t15 (t9 can't be assigned)	C5 (5) t5, t9, t11, t12, t14	
R6	C3 (1) t15	C10 (1) t3	
R7	C10 (2) t4, t5		
R8	C1 (1) t15	C10 (1) t6	
R9	C9 (2) t6,t7		
	C10 (2)		
R10	t7,t8		

In this model, the associate dean of the university and the system (responsible for assigning the centrally controlled resources and interested in optimizing welfare of faculty and students) act as a leader, in that it may hold its faculty assignment (its strategy) fixed while all other agents (the followers) react independently to the leader's strategy, reaching a Nash equilibrium relative to the leader's strategy. As in game theory, this Stackelberg games, and resulting Stackelberg equilibria of the model for the faculty assignment is induced by a strategy that is precisely the optimal assignment of all of the courses.

V. Conclusion

In conclusion, we propose an innovative method to solve the faculty assignment and timetabling problems for the university using two-steps approach. The problem is solved in a short running time of 30 minutes and it assists management team in term of avoiding tedious and manual planning. It also takes faculty preference into consideration for the course, thus the outcome sharply reduces the conflict and, thereby, improves productivity and yields higher satisfaction of the faculty members. By using this approach, we also reduce the number of variables and errors in runtime, thus it can be used for small and medium size university for their resource planning project.

The limitation of our model is as two folds. First, we assume that the demand for the course is known and it does not vary too much since the courses starts. Otherwise, it may not be economically viable to run the course if the number of students in the class is less than the breakeven number. We can resolve this during the operations by cancelling the course, if we realize the actual number of sign up is lower than the breakeven number. However, this can only be done before the start of the course. Once the course starts, even if some students drops off half-way and the number of students in the class falls below the breakeven number, the course still continues and remains sub-optimal. This is to avoid any disruption in students' study plan or delay their graduation. Second, we assume that we always have enough adjuncts or faculty members to teach a course. If there is a mismatch, the model might not find a feasible solution. This can be overcome by ensuring enough faculty to teach the course. Otherwise, the course may not be offered until we find a faculty to teach it.

This approach can also be used in other related industry where resource assignment or planning is required. In finance, we can use our model to solve the allocation of financial advisor to potential investor and scheduling appointment. In healthcare, we can deploy our model to schedule doctor and patient face-to-face appointment or operation theater scheduling based on doctor's preference time slot and operation theater's availability.

For the next step forward, we want to explore solving the problem as one-step approach. We have acknowledged that this

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

problem might be too big or complex to be solved using the mathematical model to find the optimal solution within a reasonable timeframe. Thus, we will develop some heuristics to solve the real-world problem in the near future.

REFERENCES

[1] Abdullah S, Turabieh H. "On the use of multi neighbourhood structures within a tabu-based memetic approach to university timetabling problems", Information Sciences 2012;**191**:146-68.

[2] Adewumi, A. O., Sawyerr, B. A., and Ali, M. M. "A heuristic solution to the university timetabling problem", International Journal for Computer-Aided Engineering and Software 2009, Vol. 26, No. 8, 972-984.

[3] Beligiannis GN, Moschopoulos CN, Kaperonis GP, Likothanassis SD. "Applying evolutionary computation to the school timetabling problem: The Greek case", Computers & Operations Research 2008; Vol **35**:1265-80.

[4] Boronico, J. "Quantitative modeling and technology driven departmental course scheduling", Omega, Vol. 28, Issue 3, pg. 327.

[5] Daskalaki, S., and Birbas, T. "Efficient solutions for a university timetabling problem through integer programming", European Journal of Operational Research 2015, Vol. 160, Issue 1, pg. 106-120.

[6] Daskalaki, S., Birbas, T., and Housos, E. "An integer programming formulation for a case study in university timetabling", European Journal of Operational Research 2004, Vol. 153, Issue 1, pg. 117.

[7] Derigs, U., and Jenal, O. "A GA-based decision support system for professional course scheduling at Ford Service Organisation", OR Spectrum 2005, Vol. 27, Issue 1, pg. 147.

[8] Dimapoulou, M., and Miliotis, P. "An automated university course timetabling system developed in a distributed environment: A case study", European Journal of Operational Research 2004, Vol. 153, Issue 1, pg.136.

[9] Dinkel, J. J., Mote, J., and Venkataramanan, M. "An efficient decision support system for academic course scheduling", Operations Research 1989, Vol. 37, Issue 6, 853-864.

[10] Ferland, J., and Fleurent, C. "SAPHIR: A decision support system for course scheduling", Interfaces 1994, Vol. 24, Issue 2, pg. 105.

[11] Fong CW, Asmuni H, McCollum B, McMullan P, Omatu S. "A new hybrid imperialist swarm-based optimization algorithm for university timetabling problems." Information Sciences 2014; **283**:1-21.

[12] Gunawan, A., Ng, K. M., and Poh, K. L. "A hybrid Lagrangian relaxation and simulated annealing method for the course timetabling problem", Computers and Operations Research 2012, Vol. 39, 3074-3088.

[13] Hinkin, T. R., and Thompson, G., M. "SchedulExpert: Scheduling courses in the Cornell University School of Hotel Administration", Interfaces 2002, Vol. 32, Issue 6, 45-57.

[14] Koide T. "Prototype system development for examination proctor assignment problem using mixed integer programming." Proceedings of the World Congress on Engineering and Computer Science 2014; Vol. 2:961-4.

[15] Mirrazavi, S. K., Mardle, S. J., and Tamiz, M. "A two-phase multiple objective approach to university timetabling utilising optimisation and evolutionary solution methodologies", The Journal of the Operational Research Society 2003, Vol. 54, Issue 11, 1155-1166.

[16] Onouchi O, Uchigaito T, Sasaki M. "Examination timetabling problem in universities.", Abstracts book of ORSJ spring meeting 2013; 66-7 (in Japanese).

[17] Tim Roughgarden, Stackelberg scheduling strategies, Proceedings of the thirty-third annual ACM symposium on Theory of computing, p.104-113, July 2001, Hersonissos, Greece

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Novel and Robust Retroareolar Region Detection in Mammography Images

Dr. P Pedda Sadhu Naik[1]
[1]Professor & HOD
[1]Dr. Samuel George Institute of Engineering & Technology, Markapur, AP,

Dr T Venugopal[2]
[2]Professor & HOD
[2]JNTUH College of Engineering, Jagityala, Telangana.

***Abstract* – Now a day, most of the women has been suffering from Breast cancer. For assessing and segmenting the breast cancer, Mammographic image analysis is a vital tool. Most of the researchers in the literature have shown that the Retroareolar (RA) region of the breast is an important part for detecting the cancer there by the performance of algorithms can be enhanced. However, all the conventional RA region detection algorithms failed to show the reliability and most of them are manual segmentation algorithms. Here, in this proposal we designed and implemented a novel and automated RA region detection in mammography images. Our proposed frame work has been divided into three steps to detect the RA region effectively. Simulation analysis has shown that our proposed algorithms outperform the traditional approaches.**

Keywords: Mammography, RA region, breast segmentation, Hough transformation and morphological operations

I. INTRODUCTION

The most common female cancer in the world is Breast cancer with an averaged 1.67 million new cases of cancer have been diagnosed in 2015. While the age adjusted incidence rates of breast cancer in India is lesser than the foreign countries, because of the huge population the load of breast cancer is high, about 1/3rd in urban and 1/9thin rural regions. The lack of population screening in India undoubtedly contributes to this statistic but more importantly, so do lifestyle, reproductive and dietary factors. There need to be systematic efforts at researching, preserving, and promoting those factors that "protect" Indian women from breast cancer. Typically, there are four main types of breast cancer: ductal carcinoma in situ (DCIS) where the cancer is confined within the ducts of the breast, lobular carcinoma in situ (LCIS) where the cancer is confined within the lobules or glands of the breast, invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC). IDC and ILC refer to the type of breast cancer where the tumor has spread from the ducts or lobules it originated from, respectively, into the surrounding tissue of the breast. Other less common breast cancers include medullary carcinoma, mucinous carcinoma, Paget's disease of the nipple, Phyllodes Tumor, and tubular carcinoma. Breast cancer is grouped into stages which indicate the invasiveness of the disease. There are four stages—I, II, III, IV—defined by the American Joint Committee on Cancer based on a combination of tumor size, lymph node involvement, and presence or absence of distant metastasis. There is also a more general classification: early/local stage where tumor is confined to the breast, late/regional stage where cancer has spread to the surrounding tissue or nearby lymph nodes, and advanced/distant stage where cancer has spread to other organs beside the breast. There has been a decline in breast cancer mortality rates of about 2.3% over the last decade due to improved screening techniques leading to earlier detection, increased awareness, and improved treatments.

II. RELATED WORK

Diagnostic of breast cancer in primary stages is vital for enhancing the full recovery probability and for mitigating the rate of associated mortality. In present days, breast cancer's early detection has been done by mammography screening, which is the most widely utilized, effective and low cost technique [1]. For detecting the breast cancer and their assessment can be done by computerized mammographic image analysis. In order to study parenchymal patterns [2], image-based biomarkers have been applied by various researchers on particular regions of interest (ROI). In this orbit, conventional works have exhibited that the image based biomarkers association is superior in the zone immediately behind to the nipple, namely the Retroareolar (RA) region [3].

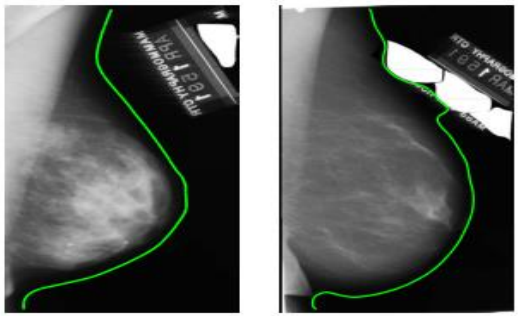


Fig.1 Example of wrong segmentation

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Later on, some amount of research has been done for the parenchymal tissue analysis in the RA region [4]. In spite of its promising outcomes, the major constraint of the said works is the requirement of human interaction for manual RA region segmentation. First most, subjectively human-composed ROIs imply certain restrictions in terms of reproducibility and scalability (e.g. the application of these methods to large datasets). Secondly, most of literature works used fixed and squared ROIs, which might be a problem for adapting shapes and sizes of wide breast verities. To address the above mentioned problems, here we proposed a novel and robust methodology for automatic RA region detection in mammographic images. We considered the input breast image geometry for automatic adjustments to diverse shapes and sizes. For this, we have built upon recent implementations in the anatomical coordinate systems creation in mammographic images [5], [6] and [7].

III. PROPOSED METHODOLOGY

Segmentation of breast is arguably the very primary pre-processing step in the algorithms of mammographic image analysis. Here, we performed the segmentation in two steps: scanning artifacts removal and contour detection. Tape artifacts are markings left by tapes, or other shadows that appear as horizontal or vertical running strips in an image. Since these are straight lines, the algorithm 2 has used to segment the foreground and for detecting the artifact lines.

Algorithm 1: Breast Segmentation
Case 1: RMLO view
Case 2: LCC view

Step 1: Select and read an input mammographic image
Step 2: Find out whether the side is left (L) or MLO using string comparison
Step 3: If the side is left then flips it horizontally
Step 4: Now, detect the breast foreground using algorithm 2
Step 5: Detect the chest walls in the case of MLO using algorithm 3
Step 6: Finally, overlay segmented mask on input breast image

For the second step, contour detection in breast, a statistical chest wall segmentation technique was used in this proposal work. By using algorithm 3 after step 6, the post processing step will be done for obtaining the binary mask by means of morphological operators in order to remove spurious artifacts and the breast contour is smoothed.

Algorithm 2: Foreground segmentation
Input: I and binary flag
where I = Grayscale mammography image with size of mxn and binary flag = 1 if the input is MLO view
Output: out_mask

Step 1: Initialize input parameters and find the central region
Step 2: Calculate maximum and minimum intensity values
Step 3: Find out the intensity threshold using the relative frequency and convolution
Step 4: Artifact removal with connected components labeling by keeping largest left-most cluster
Step 5: Now, find the contour points using region boundary tracing
Step 6: Finally, obtain the out_mask by using curvature analysis

Algorithm 3: Chest wall segmentation
Input: grayscale mammography image
Output: segmented chest wall
Step 1: First, crop the mammography image according to the contour
Step 2: Apply dilation and filtering for pre-processing of input mammography image and replace Step 3: the lower right corner with zeros
Step 4: Now, detect the pectoral line using Hough transformation then calculate the edge co-ordinates
Step 5: Calculate the accumulation array 'A' by quantizing the parameter spaces
Step 6: Find maximum in 'A'

IV. SIMULATION RESULTS

All the experiments have been done in MATLAB 2016b environment with 4GB RAM and Intel processor. The Digital Database for Screening Mammography (DDSM) [9] dataset has been utilized for the testing purpose, which consists nearly 2500 mammogram cases from various medical institutions in United States of America. Here in our experiments, we had considered the images with LCC and RMLO view by scaling them with factor 8 and later on converted to unit8 for enhancing the processing speed.

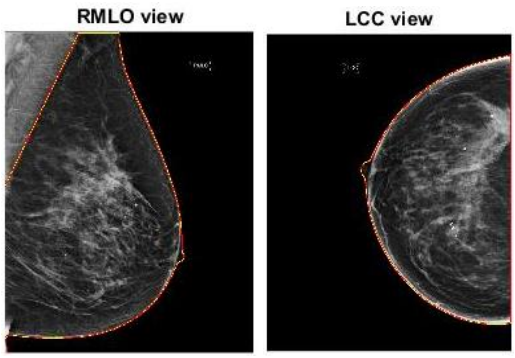


Fig.2 Segmented RMLO and LCC view mammography images

Table 1 show that the number of tested mammography images with normal and abnormal conditions. The number of correctly segmented count

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

also measured in the table with their accuracy in % values. Totally we tested 72 images and got an accuracy of 73.61% with our proposed methodology. Figure2 show that the segmented output images of RMLO and LCC view images with RA region detection, which have been obtained by utilizing our proposed algorithms discussed in section III.

Table I: Performance of breast segmentation

Type	Analyzed	Correct	Accuracy
Normal	47	35	74.44 %
Cancer	25	18	72 %
Total	72	53	73.61 %

Performance of proposed and conventional RA region detection in terms of accuracy has shown in figure 3. The relation between specificity and sensitivity shown in figure 4

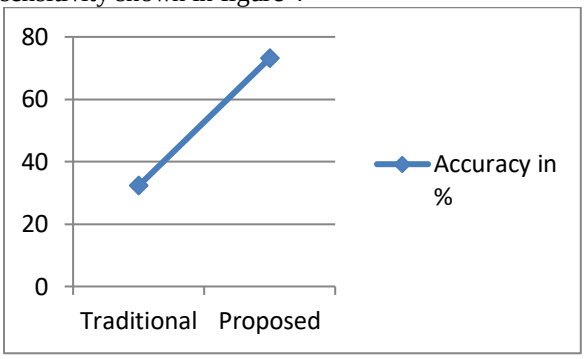


Fig.3 Performance of proposed and traditional methods

Experimental results show that the proposed method outperforms selected ROIs for the cancer detection task using proposed method.

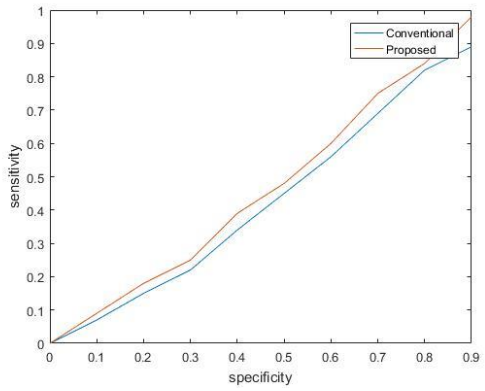


Fig. 4 Relation between specificity and sensitivity

V. CONCLUSIONS

Here, we proposed a novel and robust RA region detection in mammography images. Our proposed algorithms show the robustness of segmentation and RA region detection in terms of accuracy and sensitivity. We tested both RMLO and LCC view mammography images for showing the effectiveness of proposed algorithm. Furthermore, this can be enhanced by applying contrast enhancement approaches for better segmentation accuracy. We can also develop RA region detection in 3D mammography images.

REFERENCES

[1] American Cancer Society. Breast Cancer Facts and Figures 2015-2016 Atlanta:American Cancer Society, Inc. 2015.

[2] Tahoces P. G., Correa J., Soutos M., Gomez L., Vidal J. J..Computer-assisted diagnosis: the classification of mammographic breast parenchymal patterns Physics in Medicine and Biology. 1995;40:103.

[3] Li Hui, Giger Maryellen L., Huo Zhimin, et al. Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: Effect of ROI size and location Medical Physics. 2004;31:549– 555.

[4] Wei Jun, Chan Heang-Ping, Wu Yi-Ta, et al. Association of computerized mammographic parenchymal pattern measure with breast cancer risk: a pilot case-control study Radiology. 2011;260:42–49.

[5] Brandt S. S., Karemore G., Karssemeijer N., Nielsen M.. An Anatomically Oriented Breast Coordinate System for Mammogram Analysis IEEE Transactions on Medical Imaging. 2011;30:1841–1851.

[6] Pertuz S., Julia C., Puig D.. A Novel Mammography Image Representation Framework with Application to Image Registration in 2014 22nd International Conference on Pattern Recognition (ICPR):3292–3297 2014.

[7] Abdel-Nasser Mohamed, Moreno Antonio, Puig Domenec. Temporal mammogram image registration using optimized curvilinear coordinates Computer Methods and Programs in Biomedicine. 2016;127:1–14.

[8] Heath Michael, Bowyer Kevin, Kopans Daniel, Moore Richard, Kegelmeyer W Philip. The digital database for screening mammography in Proceedings of the 5th international workshop on digital mammography:212–218Citeseer 2000.

AUTHORS PROFILE

Dr. P Pedda Sadhu Naik[1] received B.Tech (CSE) Degree from JNTUniversity in 2003 and M.Tech (CS) Degree from JNTUCE Anatapur in 2007.He is received Ph D from JNTUK, Kakinada in Computer Science under the guidance of Dr T. Venu Gopal. He has 14 years of teaching experience. He joined as Assistant Professor in Dr.Samuel George Institute of Engineering & Technology, Markapur, India in 2003. Presently he is working as Associate Professor and Head of the Department of CSE. His Interested research areas are Image Processing and Soft Computing. He has life member of ISTE and IAENG. He organized various National level Technical Symposiums as Convener. He attended Various National and International Workshops and Conferences on image Processing.

Dr T Venugopal[2] received B.Tech (ECE) Degree from Osmania University Hyderabad in 1994 and M.Tech (CSE) Degree from JNTUH Hyderabad in 2003.He completed Ph D from JNTUH, Hyderabad in Computer Science. He is Professor, JNTUH College of Engineering, Jagityal. Jagityala District, Telangana. He has 20 years of teaching experience. He has worked in different colleges in different

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

levels. Presently he is working as Professor in JNTUHCE Jagityala. His Interested research areas are Image Processing and Soft Computing. He has life member of ISTE and IAENG. He published more than 50 international journals. He organized various National level Technical Symposiums as Convener. He attended Various National and International Workshops and Conferences on image Processing

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Improving DBMS Security through the use of a Checksum Technique

Naim Ajlouni* Firas Ajlouni** Alaa Ali Hameed***

Istanbul Aydin University* ***
Lancashire College of Further Education**
naimajlouni@aydin.edu.tr

Abstract
We propose an approach to database security that exploits existing DBMS facilities to associate a separately maintained checksum value with critical data. Using our approach, a database's content, domain and referential integrity remain the responsibility of the DBMS, however, when critical data is manipulated checksum values are computed and stored in a separate database. Using this combination of databases, applications which access critical data can only access such data via checksum values ensuring that data is created and accessed in a secure manner.

Key words- DBMS, Security, Checksum, entity integrity, domain integrity, referential integrity and user-defined integrity.

1. Introduction

The primary responsibility for ensuring the security and integrity of a database lies with the **D**ata**b**ase **M**anagement **S**ystem (DBMS) [1]. At present, the most widely used **R**elational DBMS's (RDBMS's) provide several ways of ensuring data integrity at the entity, domain, referential, and user-defined levels. The latter of these levels, i.e. the user-defined level, involves constraints on the forms of update that can be performed, and such constraints are usually enforced via triggers either by the database itself or by applications that access the database.

Checksums have been the subject of research and practical application for many years. Gopalan [5] used checksums to enhance the integrity of a conventional file system via a block-level checksum computed per-block for all data blocks in a file, which is indexed by a relative block number. In such an approach, files are modified to include a set of additional references that point to the checksum blocks, and the addition of a block to a file involves the computation and storage of its checksum blocks. Crocker [6] introduced a checksum-based "Spam Detection" engine for Lotus Domino mail systems called *Block* which also exploits checksums. The system computes a checksum for each message that is classified as "Spam" and maintains the checksums in a database which is replicated between the mail system and the mail server. The database enables the server, on receipt of each message, to compute a checksum and compare it with those checksums already in the database. If a match is found, it is likely that the associated message is further "Spam". Network Appliance Inc. technical report [7] describes a technique for reducing the volume of data transfer during Backup and Restore on UNIX platforms. The system uses a checksum to identify portions of a file that have changed since a previous and current backup. Changed blocks are identified via checksum values computed and maintained for each block. Sabartnam [8], uses checksums to detect errors in database manipulations including transactions, locks, logs, and data buffers. In this case, a checksum is added to the object being manipulated, and a checksum field is attached to the access method and also to each of the objects in associated hash buckets. A hash vector and object checksum are recomputed during the restart of a DBMS and compared with stored checksum values.

We describe an approach to database security that similarly exploits checksums. First, in Section 2, the forms of data integrity that must be maintained in a conventional RDBMS are examined. In Section 3 we present our approach to maintaining user-defined integrity using error detection and correction algorithms similar to those implemented in network protocols, i.e. the Hamming Code Protocol. [3] [4]. An implementation of the proposed approach is explained in detail in Section 4, Section 5 test and Section 6 examines properties of the implementation.

2. Data Integrity in an RDBMS

Data Integrity levels in an RDBMS can be classified into four main kinds, i.e. entity integrity, domain integrity, referential integrity and user-defined integrity [2].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Entity Integrity: Useful at the row level in a table, entity integrity ensures that a relation does not have any duplicate rows and that each

- have a unique primary key that can be defined by one or more of its attributes.

- **Domain integrity**: Values in a column in a table must be drawn from some well-defined "domain" of values. This is the simplest form of an integrity constraint which is maintained at all times and in all circumstances. In effect, a "domain" corresponds to *type* in a conventional programming language, and values in a column must be drawn from one of the available types.

- **Referential Integrity:** This is applied at the table level such that values available in one relation are available and synchronized with those in other relations. Referential integrity is enforced with a primary key and ***F****oreign* ***K****ey* (FK) combination. A foreign key comprises one or more columns in a "child relation" whose values are synchronized with those in the PK in a "parent relation". The FK accepts only those values that exist in the PK in order to maintain the integrity of the relation. Referential integrity is preserved when applying any ***D****ata* ***M****anipulation* ***L****anguage* (DML) operations, i.e. insert, update and delete operations, via the following constraints on the application of such operations:-

  1. Restricted: Disallow data modification.
  2. Cascaded: Extend the data modification on parent relation to all child relations.
  3. Nullified: Set the values of matching FK's in child relation's to the value NULL.

- **User-Defined Integrity:** In effect, user-defined constraints on the ways in which a database can be manipulated. Such constraints are the responsibility of the system administrator, who will administer access rights and enforce rules and regulations.

The technique described in the following section is intended to deal with user-defined integrity, a form of integrity typically enforced via triggers and constraints [9]. Our technique is intended to augment conventional approaches to user-defined integrity by exploiting conventional triggers and constraints to ensure that access to critical data is only possible via checksums.

3. Database Checksum Requirements and Prerequisites

There are several requirements that our approach must satisfy:-

- As the potential exists for "illegal" data modifications to change all the stored data, any solution should be able to both detect and correct all the modifications to data in such circumstances (completeness)

row in a relation has its own unique identifier which distinguishes it from other rows. This unique identifier is termed the ***P****rimary* ***K****ey (PK)*, and each relation must

- The solution should not be restricted in the domains whose values may be used in the computation of the checksums (no special cases)
- The solution should involve no additional overhead where legal modifications are made (efficiency)
- The solution has the prerequisite that both the original database and its associated checksum database are available to all applications which access the original database, and that all such applications correctly maintain both the original database and the associated database of checksums (correctness)
- Execution time overheads and additional space requirements should both be minimal (efficiency)

3.1. Our Approach

Our approach involves the use of two databases, the first database contains conventional data of some kind and the second stores associated checksum values, i.e. a result and a remainder, that are associated with critical data in the first database. Figure (1) shows to example databases named "DB1" and "CS1, "DB1" is a conventional database is containing values of any permissible type, e.g. integer, decimal, date including time, etc. The checksum values for critical data in "DB1" are stored in "CS1".

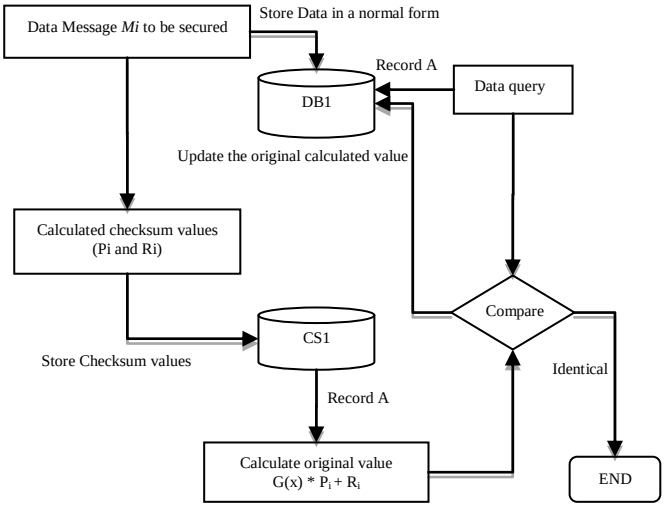


Figure (1)

Checksum values are calculated as the modulus value of a data item (the item being first converted to its ASCII value if necessary) using a divisor chosen by the system administrator.

The equation for the recalculation of the original values using the stored checksum related values:

$$D_i = G(x) * P_i + R_i. \qquad 1$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

where D_i is the recalculated value, $G(X)$ is a fixed regular divisor chosen, e.g. 1028, P_i is the result of the division of M_i by $G(x)$, and R_i is the modulus of M_i and $G(x)$ are given by,

$$P_i = M_i / G(x) \quad 2$$

The modulus value is calculated using equation (3) as shown bellow:

$$R_i = M_i - (G(X)*(M_i \setminus G(X))) \quad 3$$

Where, M_i is assumed the i^{th} datum in a database of N data sets, \ is the remainder, in this case P_i and R_i are the checksum values for the i^{th} datum in a database of N sets of data, were (N is $1,2,3,\ldots\ldots N$), i.e.

By substituting equations (2) and (3) into equation (1) we yield

$$D_i = G(x) * M_i / G(x) + M_i - (G(X)*(Mi \setminus G(X))) \quad (4)$$

Equation (4) can be further simplified to give,

$$D_i = 2M_i - (G(X)*(Mi \setminus G(X))) \quad (5)$$

The checksum values P_i and R_i are maintained separately, e.g. in the database called "CS1" in Fig. 1., and used later to validate the integrity of the i^{th} data item stored in the original "DB1" database. When a data item in the "DB1" database is to be accessed by an application, the application must a) compute checksum values, and b) compare the computed values with those stored in the "DB1" database. If the computed and stored checksum values differ, then the associated critical data in "DB1" database is assumed to have been changed illegally, and the stored checksum values for the i^{th} data item are used to recalculate the original values and restore the database to a consistent state.

Consider, next, the worked example below:-

i. Integer Test

Let M_i denote a "critical" credit card number that is to be inserted in a "DB1" database of account information, and $G(x)$ denote a fixed regular divisor 1028. For simplicity, we assume that the initial balance in all accounts is £0.00. Table (1) bellow, shows a table of stored values in the Data base with their related Check sum values.

M_i	G(X)	P_i	R_i	D_i
7654321678945320	1028	7445838209099	548	7654321678945320

Table (1) stored data

Keeping in mind that the checksum values for P_i and R_i are stored in the "CS1" database and the value M_i is stored in "DB1" database together with other account data including the opening balance.

Assuming that subsequent to the initial insertion of the account numbered 7654321678954320 into the "DB1" database, an illegal update takes place and the credit card number is changed to (7654321678955555). Table (2) bellow shows the content of the data base after the illegal modification has taken place.

M_i	G(X)	P_i	R_i	D_i
7654321678955555	1028	7445838209099	548	7654321678954320

Table with illegally changed M_i Value (2)

Any query issued on this account (or any other critical data) will lead to an automatically re-computing the checksum values using equation (5) which will be followed by a comparison of both recomputed and stored valued of the checksum, comparing the computed values with the corresponding stored value as shown in Figure (2). Which lead to a restoration of the original value in the database to preserve database integrity?

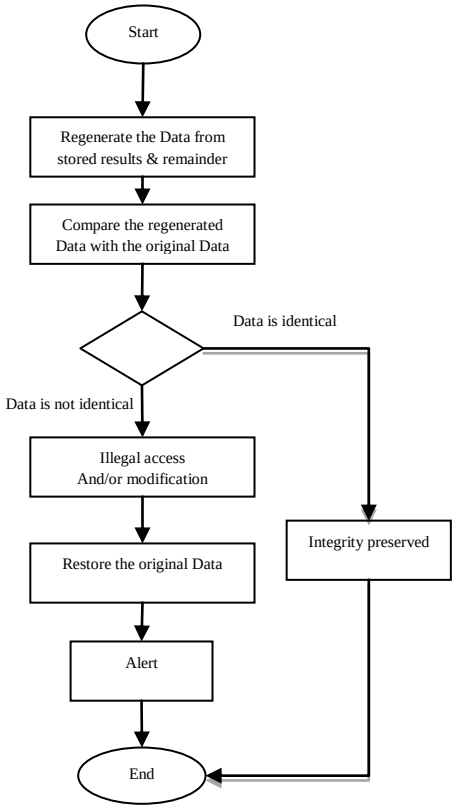


Figure (2)

If $M_i = D_i$ then the critical data is assumed to be valid, otherwise the critical data is assumed to be incorrect, but can be recovered provided only that the values of the data involved were inserted by an authorized user.

ii. Character Test

Assuming that we have a string M_i whose value is "Rob" that is to be inserted into a "DB1"database whose fixed divisor, G(x), is 1028. The string is first converted to a sequence of values that correspond to its character's ASCII codes. However, all such ASCII values have the value 100

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

added to ensure that they are 3 digits in length. The resulting ASCII values are concatenated together to obtain a single number.

Assuming that M_i = "**Rob"**, the ASCII equivalent is 8211198, i.e. ASCII for **R** is 82, for **o** is 111 and for **b** is 98 giving 182 211 198 when the necessary 100 is added.

If G(x) is 1028, the values for P_i and R_i are shown in table (3) bellow

Character Set	M_i	G(X)	P_i	R_i	D_i
Rob	182 211 198	1028	177248	254	182 211 198

Table (3) stored data

Assuming that subsequent to the initial insertion, an illegal update occurred and the name was changed to "John". Any query issued on this critical data requires the checksum values to be recomputed and compared with the stored values, i.e. the value associated with the initial insertion of "Rob", using equation (5).

Each 3 digit substring of D_i represents a character, which, when 100 is subtracted from the three digit substring yields the ASCII codes 82, 111 and 98 enabling the value in the "DB1" database to be recovered.

4. An implementation of the proposed solution

For simplicity, it is to be assumed that all access to data in a "DB1" database is protected through the use of an *application view* [12], and direct access is forbidden by the DBMS's access policy which is set by the database administrator and subject to any other third-party security system, e.g. a Firewall [10, 11].

Application triggers then perform the calculation of checksum values, i.e. result and remainder P_i and R_i respectively and their storage in the "CS1" database.

Consider, next, a customer payment system, composed of several tables, among which are the three tables shown in Figure (3) below.

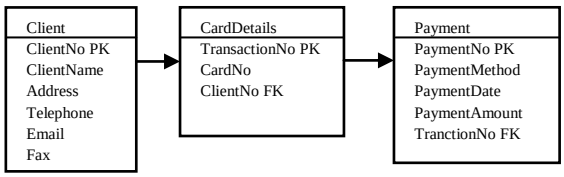


Figure (3) Case Presentation

Our interest here is to preserve the integrity and security of the following fields:

- PaymentAmount Field in the Payment table.
- CardNo Field in the CardDetails table.
- ClientName field in the Client table.

As suggested earlier, direct integrity rules provide basic integrity protection, but are unable to prevent illegal changes to values stored in a database, and similarly unable to prevent the illegal exposure of such values. In such circumstances database transactions, both "legal" and "illegal" will run normally and it will be assumed (even in the event of an illegal transaction) that the integrity of the database is preserved and uncompromised.

4.1 Validation Process

The validation process is concerned with data integrity; this process is carried out by recalculating the data related to the data checksum value using a pre saved indicators related to this checksum, and comparing those values with available data values. The validation algorithm is shown below:

```
Begin
        Identify record in Client table;
        Obtain CardNo value for defined record;
        Obtain result, remainder values for
                                    defined record;

        Compute Checksum
            Checksum = result X G(x) + Remainder
        Convert checksum to string;
        if checksum = 0 and
           (CardNo value = 0 or CardNo = null)
          { validate = true;}
        else if checksum ≠ 0 and CardNo =null
          { validate = false;}
        else
         {if checksum = CardNo
              {validate = true;}
              Else
             {validate = false;}
        }
End
```

4.2. Number implementation

This example will be applied to the CardDetails and NCalculations as shown in Figure (4):

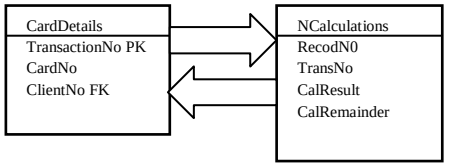


Figure (4) Presented Case

CardDetails Table: Any illegal changes to values of the CardNo in this table must be detected and restored.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

NCalculations Table: Used to maintain the checksum values for corresponding CardNo values.
Our approach allows users to perform basic data manipulation operations such as Insert, Update, Delete and Select. Two types of triggers are then used, i.e. conventional triggers [9] and block-type triggers [9].

Conventional Triggers

- **Pre-Query Trigger. U**sed to delete any data inserted illegally in "DB1" database by unauthorized user, and to check user authorization, see Figure (5).

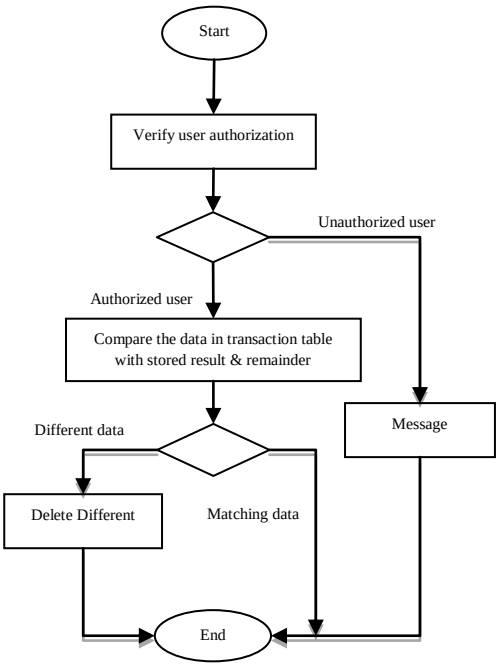


Figure (5)

- **Pre-Insert, Pre-Update & Pre-Delete Triggers.** These triggers check user authorization, and if the user is unauthorised an error message is generated. See Figure (6)

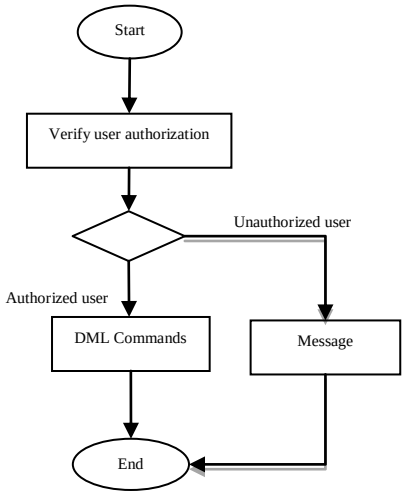


Figure (6)

Block Triggers

- **Pre-Insert Trigger:** This trigger calculates the checksum values i.e., P_i and R_i using equations 1 and 2, and stores the values in the NCalculations table in the "CS1" database. See Figure (7)
-

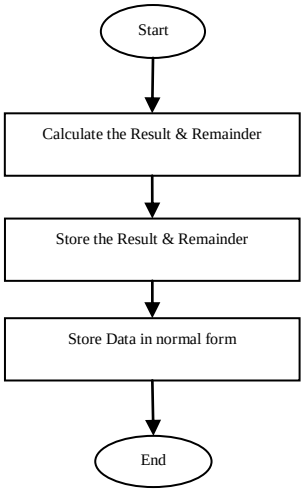


Figure (7)

- **Post-Query Trigger:** This trigger is used after any query to re-compute the stored checksum values using equation (3) to regenerate the original values stored in the "CS1" database. Recomputed values are compared with the values stored in the transactional table in "DB1". If they are not the same, the recomputed value will is stored in "DB1". See Figure (8)

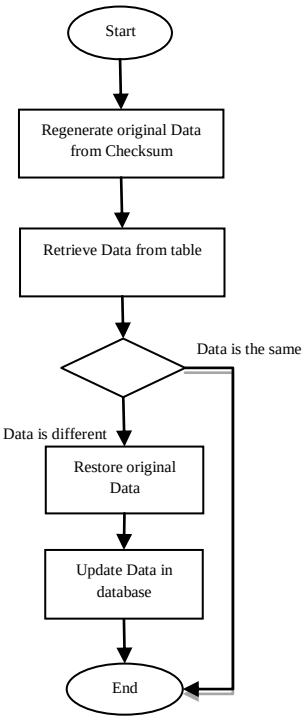


Figure (8)

- **Pre-Update Trigger:** This trigger provides the means to update the values record in "DB1" and "CS1".

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Updates are forced on the Ncalculations table which contains the values P_i and R_i. The update uses the recalculated checksum values i.e. the P_i and R_i. See Figure (9).

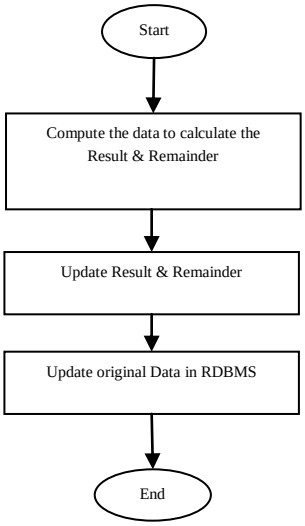


Figure (9)

- **Pre-Delete Trigger. This** trigger deletes any values from the transactional table in the "DB1" database and the related checksum values in the "CS1" database, i.e. delete the P_i and R_i

4.3 Character implementation

This implementation example involves the tables in Figure (10). The **Client Table** contains the client transactions information. Illegal changes to the ClientName in this table must be detected and restored. The **Calculations Table** stores the checksum values of P_i and R_i , i.e. the result and remainder.

The implementation must provide users with a means of performing the basic DML operations Insert, Update, Delete and Select. Two types of triggers are used, i.e. f Conventional triggers [9] and block triggers [9] of the same kind used the implementation in Figures 5 and 6.

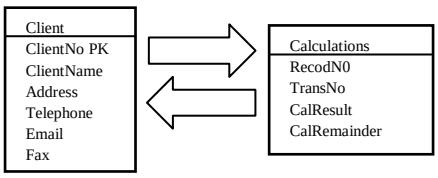


Figure (10)

Block Triggers

- **Pre-Insert Trigger:** This trigger converts the text ASCII codes and calculates the checksum values P_i and R_i which are then stored in the Calculations table in the "CS1" database. See Figure (11) below.

- **Pre-Update Trigger:**

This trigger provides the means to update the values in the "DB1" database as well as the checksum values in the "CS1" database. Updates are forced on the Calculations table containing the checksum values P_i and R_i. See Figure (12). This update uses the recalculated checksum values for P_i and R_i.

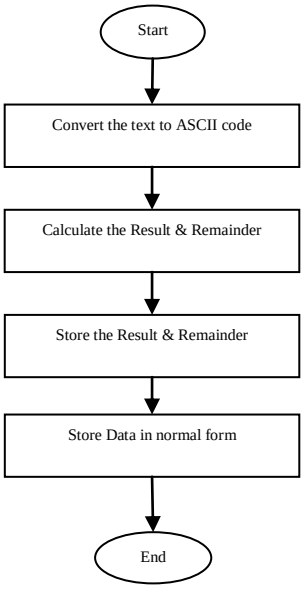


Figure (11)

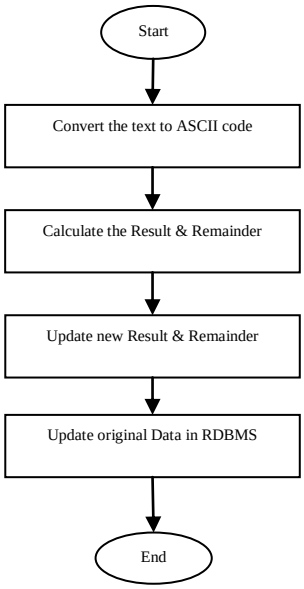


Figure **(**12)

- **Post-Query Trigger:** This trigger is used after all queries and re-computes the stored checksum values using equation (3), i.e. it re-computes the ASCII values and then converts the ASCII codes to a string. The checksum values in "CS1" are then compared with the data stored in the transactional table in "DB1" and, if they are unequal, the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

correct value is automatically inserted into “DB1”. See Figure (13).

Using our approach, conventional user authorization is maintained via a Username & Password combination which must be provided when connecting to a database. In addition *Roles Assignment* [9] ensures that authorized users are granted the required role and that their transactions are thus valid. In our examples, users were assigned the role “CHECKSUM” in order that they are able to make legal changes.

5.0 Test and Results

As a simple test to determine the correctness of the proposed approach, 11 random records have been entered into a “DB1” database. The test data was then updated by illegal users using direct access to the “DB1”.database through SQLPLUS where all ClientName and CardNo attributes where updated and committed as shown below.

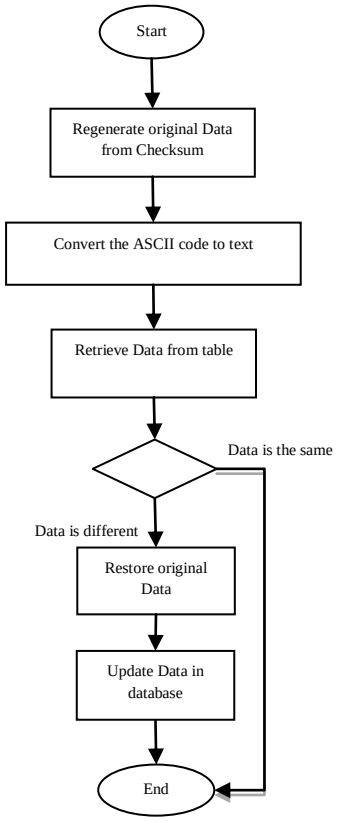


Figure **(**13)

5.1 Character Test

The following SQLPLUS command was used to check the client name from the table called client.

SQL> select clientname from client;

The returned result as displayed below:-

CLIENTNAME

Allen Marlard
Ward Jones
Martin King
Adams Blake
Martin Jones
Scott Clark
King
Turner King
Adams
James Miller
Miller King

11 rows selected.

Assuming the client names in the table client are to be changed illegally by directly connecting to the SQL to “DB1”database, the illegal update to be executed and committed requires the following commands:

SQL>update client set clientname =' Oliver';
11 rows updated.

SQL>commit;

commit complete.

If we recheck the constants in the table client in the “DB1”database, all client names have been changed to OLIVER as shown below.

SQL>select clientname from client;

CLIENTNAME

Oliver
Oliver
Oliver
Oliver
Oliver
Oliver
Oliver
Oliver
Oliver
Oliver
Oliver

11 rows selected.

It should be mentioned here that the critical data in the “DB1” database at each query by the proposed application is checked by re-computing the checksum values and comparing them with the data stored in the “DB1”database. If the data values are equal the content is considered to be OK, however if they differ the original values are restored using the computed checksum values. Assuming that a legitimate user queries the “DB1” database using our application, and if at the same time checks are made on the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

content of the table *client* it can be seen that the original data has been restored.

We select the ClientName again from SQLPLUS as below.

SQL> select clientname from client;

CLIENTNAME

Allen Marlard
Ward Jones
Martin King
Adams Blake
Martin Jones
Scott Clark
King
Turner King
Adams
James Miller
Miller King

11 rows selected.

This test determines if an illegal user inserts a record into the "DB1"database, and, if so, the illegal insertion is removed from the "DB1" database as follows.

First the record is inserted and committed:-

SQL> insert into client values (100,'CATHIE','88 Park Road',
'0161383621','Cathie@hotmail.com','0161383620',1);

1 row created.

SQL> commit;

Commit complete.

A recheck of the table *client* then indicates that *Cathie* has been added in the as shown below.

SQL> select clientname from client;

CLIENTNAME

Allen Marlard
Ward Jones
Martin King
Adams Blake
Martin Jones
Scott Clark
King
Turner King
Adams
James Miller
Miller King
Cathie
12 rows selected.

Queries on the "DB1" database using our application forms involve a check on the content of the table *client*, and it can be seen that the original data has been restored and the illegal record has been removed, i.e. selecting the *clients* as shown below:

QL> select clientname from client;

CLIENTNAME

Allen Marlard
Ward Jones
Martin King
Adams Blake
Martin Jones
Scott Clark
King
Turner King
Adams
James Miller
Miller King

11 rows selected.

5.2 Number Test

The same type of test was performed on the CardDetails table. The results were the same as the character test. In the first test, the CardNo was updated by an illegal user and it successfully restored, in the second test a new record was illegally inserted, then detected and removed.

The two tests have been carried out on an implementation of out model. In both tests the system suffered from illegal access via direct access to the "DB1" database where several records were updated and/or new records inserted in the Client and CardDetails tables.
The test results indicate that the model is working, and the performance of the system was monitored during tests and was not adversely affected by the overhead of computing and using checksum values.

Table (4), shows the actual processing times, for the update comparison operation on the test database, the test was carried out using different database sizes. As it can be seen that the data integrity recalculation process carried out affects the performance of the system, however this effect is within the acceptable time margins and would not affect user during a normal run of the system as the delay is negligible.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

No. of Records	Scenario1 regular DB Milliseconds	Scenario2 DB with Checksum Milliseconds	Checksum Overhead
10,000	31,236	68,406	2.19
20,000	62,187	137,813	2.22
30,000	94,000	203,765	2,17
40,000	127,406	271,765	2.13
50,000	158,094	340,219	2.15
Average			2.17

Table (4) Test Results

Figure (14) shows the actual test data representation of both scenarios the first being without checksum parameters and the second scenario with the checksum parameters.

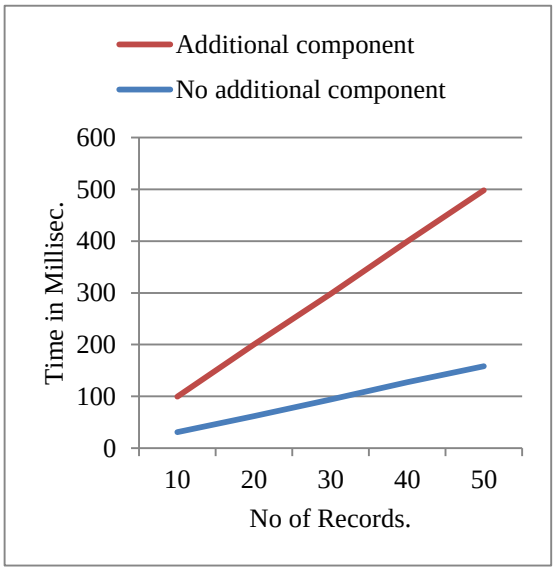


Fig (14) Oracle 10G Checksum Update overhead

6. Conclusions

We have presented an approach that can be used for all types of data; this approach is based on using a checksum validation algorithms applied only to critical data such as credit card numbers. This approach protects the data from any illegal changes and guarantees data integrity. In this case the checksum values can be stored in the same database as the regular data or it can be stored in a separate database i.e. “CS1” database, this will increase the level of security. It was also shown that the algorithm provide the necessary data integrity needed for any system regardless of it size and or data type. The proposed data checksum mechanism has an advantage that it can detect whether data has been modified, and in this case the algorithm compute the original data, this will provide a data integrity mechanism to protect the data.

Future work can be implemented based on this algorithm were the integrity of encrypted data can be maintained, in this case the checksum values can be calculated for the encrypted data, the checksum values are then stored to be used at any time to reconfirm the encrypted data validity. This will protect encrypted data from any illegal changes that might take place by unauthorized users. Therefore this algorithm is capable of protecting any data type from illegal access and or changes, especially changes such as replacing data by similar type but different in value into the database.

References

1. Jingmin He and Ming Wang. Cryptography and Relational Database Management Systems. Procedings of the International Database Engineering and Application Symposium (IDEAS'01), 2001, IEEE, Computer Society.
2. Silberschatz, Korth, Sudarshan. Database System Concepts, 4th edition.
3. Dorothy E. Denning, Cryptographic Checksums For Multilevel Database Security
4. Andrew S. Tanenbaum, Computer Networks, 4th edition. Prentice Hall. 3rd chapter.
5. Gopalan Sivathnu, Charles P. Wright, and Erez Zadok, "Enhancing File System Integrity Through Checksum", Technical Report FSL-04-04, Stony Brook University.
6. Darren R. Crocker, 2004, Figurehting Spam on Lotus Notes/Domino using no.Spam.domino, Block Checksum detection, Whitepaper.
7. Darrin Chapman and Toby Creek. "Oracle and Open Systems SnapVault Backup and Restore for UNIX Platform".Network Appliance Inc. TR 3377, 2005.
8. Maitrayi Sabaratnam, "Cost of Ensuring Safety in Distributed Database Management Systems". Norwegian University of Science & Technology.
9. Oracle8i Application Developer's Guide - Fundamentals. http://www.csee.umbc.edu/help/oracle8/server.815/a68003/toc.htm
10. Oracle Database Security Checklist, An Oracle White Paper January 2006
11. The Database Holds Your Core Assets—Protect It First, http://www.devx.com/dbzone/Article/11961/1763

12 Database views, http://publib.boulder.ibm.com/infocenter/wbcric/v251/index.jsp?topic=/com.ibm.wbcr.doc/apidoc/i_bcr_r_apidatawbcr.html

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Coupling GIS-based MCA and AHP techniques for Hospital Site Selection

Aseel Kmail
Jacqueleen Jubran
Walid Sabbah
Geographic Information System Department, AAUJ
Jenin, Palestine
aseel.kmail@aauj.edu
jacqueleen.joubran@aauj.edu
walid.sabbah@aauj.edu

Abstract

Recently, the population of Jenin city is increasing rapidly and this amplifies the need for more infrastructural objects such as hospitals. Hospitals are considered among the most important infrastructural constructions in cites as they provide health care services. However, current hospitals and medical resources are limited and randomly allocated in Jenin city. Accordingly, in this paper we propose a suitability model that employs Geographic Information System (GIS) based Multi-Criteria Analysis (MCA) with Analytical Hierarchy Process (AHP) to identify suitable locations for building a new hospital in Jenin city. An experimental instantiation of the proposed model is instantiated and the produced results show that the majority of suitable areas are located in Northeast of Jenin. This is mainly because Northeast of Jenin is away from industrial areas and dumping sites.

Keywords: *Hospital site selection, GIS, MCA, AHP*.

I. Introduction

Recently, the number of people shifting from rural to urban areas is increasing rapidly. This rapid increase of urban population creates various social, economic and environmental changes such as unplanned sprawl, inadequate housing facilities, traffic congestion, insufficient drainage, and lack of health facilities [1]. It is very important to provide all facilities and infrastructural constructions in urbanized areas to overcome the rapid urban growth. Therefore, it becomes the government's responsibility to provide the required resources and facilities for urban areas on proper locations.

Hospitals are among the most important facilities that have a vital role in providing health care services. Identifying the best locations for new hospitals is an important issue due to the fact that selecting suitable locations will help the government to optimize the allocation of medical resources, simplify social contradictions and control the health care development in rural and urban areas. On the other hand, appropriate hospital site selection will help people reach hospitals easily, reduce the time of rescue and improve the quality of life [2].

Current hospitals in Jenin city are randomly distributed and arbitrarily allocated due to the unreasonably distribution by the government. For example, inside the city center of Jenin, hospitals are saturated. With the growth and extension of Jenin, the population increases rapidly and spreads into areas outside the city center and the contradiction between supply and demand for hospitals is becoming severe. Moreover, there is a persistent need to build quality hospitals that provide professional health care services due to the limited high quality medical buildings.

Several studies employed GIS techniques and products to address the problem of identifying the best locations for building hospitals and planning health services [3, 4]. Most of these studies take into account several parameters to allocate suitable sites for building hospitals such as existing hospitals, population, economical factors, pollution, and other laws and regulations.

In this paper, we aim to identify the most suitable areas for building a new hospital in Jenin city. In order to achieve this goal, we will exploit GIS products and methods with MCA in addition to AHP. By this, we mean that the study will take into account many factors such as existing hospitals, proximity to main roads, and distance to polluted and industrial areas. After that, we will assign them different weights (according to their importance) based on AHP.

The main contributions of our work are summarized as follows:

- Employing GIS-based MCA in order to identify the best locations for building a new hospital in Jenin city.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Exploiting AHP to assign weights and scores for the identified criteria (i.e. factors) in order to select the best location for the new hospital.

The rest of this paper is organized as follows. Section 2 presents the related work. A general overview of the study area is presented in section 3. Section 4 presents the general architecture of the proposed model and the implementation details of the suitability model. Experimental validation and evaluation of the proposed model is presented in Section 5. In Section 6, we draw the conclusions and outline future work.

II. Related work

We will clarify our contributions in the following paragraphs by offsetting them with prior related work. Several studies have employed GIS techniques and methods in health services and for planning public health [3, 4 and 5]. For example, the authors in [6] combined GIS with Location Based Services (LBS] in order to settle the affairs of emergency medical incidents. On the other hand, other authors employed GIS techniques and methods in selecting the best site for building health care facilities. In order to build constructions that provide health care facilities, various parameters (i.e. factors) can be considered to identify the most suitable sites such as existing health care facilities, population, economic factors and pollution. These parameters can be classified, analyzed and integrated together in different methods. For example, MCA is used to identify factors that affect building new health care objects in [7]. While in [8], the researchers employed both GIS and Analytical Hierarchy Process (AHP) to determine the parameters that affect the physical accessibility of neurosurgical emergency hospitals in Sapporo city. At the same time, the authors of [9] exploited AHP to evaluate the appropriateness of the location selected for Taiwanese hospital.

Although AHP allows multi-criteria decision-making, it suffers from the fact that there are hidden assumptions like consistency. Besides, it is difficult to use when there is large number of criteria. To overcome these problems, Fuzzy Analytical Hierarchy Process (FAHP) is used later in for hospital site selection [10].

In our proposed work, we aim to employ MCA based on GIS methods and techniques to identify the best site for building a new hospital in Jenin city. Besides, we will exploit AHP to assign weights for the factors that affect the new hospital site selection. According to the produced results, we can prove that GIS-based methods and tools play a vital role in making effective decisions in the health field.

III. Study Area

The study area is Jenin governorate. It is located in the north of West Bank as shown in Figure 1. In 2016, the city had a population of 318,958 according to the census by the Palestinian Central Bureau of Statistics [11]. It is located about 43 Kms north of Nablus, and it is about (100-250 m) above sea level. The name of Jenin was derived from Ein Ganim meaning "the spring of Ganim" and referring to the region's plentiful spring.

Jenin is under the administration of the Palestinian Authority. Today, Jenin is built on the slopes of a hill and surrounded with different types of trees such as carob, fig, and palm trees. It is distinguished by its agriculture, producing various types of crops.

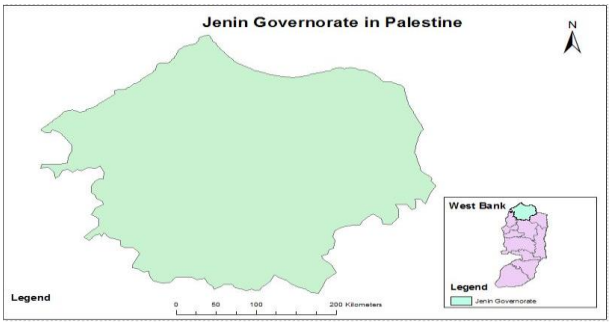


Fig.1: Jenin governorate in West Bank

Jenin governorate has 82 localities and one camp, and we divided the study area into three main regions: Jenin city, Jenin camp and villages that belong to Jenin governorate.

During our work, we focus on Jenin city that has three main hospitals. The details of these hospitals are given in Table 1, and their locations are illustrated in Figure 2.

Table 1: Existing hospitals in Jenin city.

Name	Specialization	Administrated by	No. of beds
Al-Razi Hospital	General	Private sector	60
Al-Amal Hospital	General	Private sector	20
Jenin Governmental Hospital	General	Government	120

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

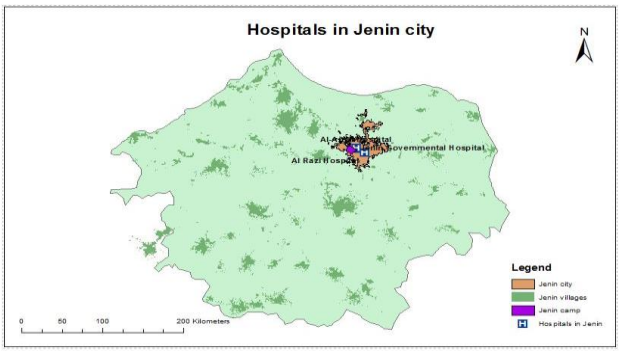


Fig.2: The main hospitals in Jenin governorate.

IV. Data and Methodology

In this section, we present the methodology used in our proposed model in order to identify the optimal site for building a new hospital in Jenin city. Figure 3 depicts the steps of our proposed model.

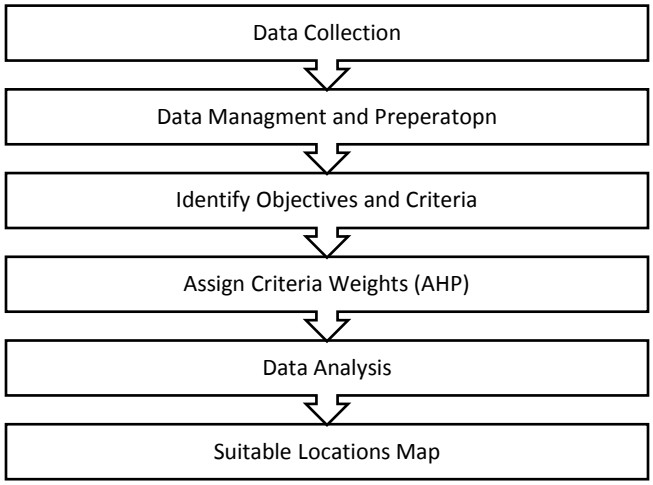


Fig.3: A flow chart depicts the methodology used in order to identify suitable locations for building a new hospital.

As shown in figure 3, the proposed model starts with collecting data from various resources to represent the aspects of the study area. Some of those data are collected from online resources such as GeoMOLG [12] and others are manually digitized with a suitable scale. After that, the collected data are managed and prepared for use. By this, we mean that the data are organized and stored efficiently for further analysis. Then, analysis objectives and criteria are identified to use them in further steps. Once analysis criteria are identified, they are assigned weights based on AHP. Those weights are used to indicate the importance of each criterion. After that, during the data analysis step, various GIS tools and methods are employed in order to produce the set of suitable locations for building a new hospital.

A. Data Collection

In order to find the best location for building a new hospital, we need to collect data in the format of vector shape files. These shapefiles are collected from GeoMOLG [12] and some of them created by digitizing maps obtained from Jenin municipality. Once the vector data are collected, they are converted later during the analysis step into raster format.

B. Data Managment and Preperation

During this step, the collected data are prepared to be used in the analysis process. Data are often collected with missing values and errors, so we need to correct these errors and organize the data in datasets and geodatabases. The process of correcting the data and integrating them into feature data sets constitute a vital role in this step.

Additionally, it is important in this step to answer some questions about the collected data such as:

- What is the data format?
- At what scale it was collected?
- Are the data projected?
- Does the data have all the needed attributes?
- Does the data have constraints and the features geometry support the analysis process?

C. Identify Objectives and Criteria

In our proposed model, we aim to select the optimal site for building a new hospital. Various factors have been involved in the selection process including the following:

1) *Land use.*
2) *Distance to existing hospitals.*
3) *Intersection with main roads.*
4) *Distance to dumping sites.*
5) *Distance industrial areas.*
6) *Elevation.*

These factors are divided into three main types:

1- Technical factors: these factors have a clear impact on the construction process and they include the elevation, the slope, distance to existing hospitals and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the land use of the proposed site. The land use refers to how the land being used by human. While the distance to existing hospitals is how fare the new hospital from other hospitals in the same city.

2- Environmental factors: there is a strong relationship between human and the environment. The main environmental concerns that may affect hospital site selection are noise and pollution. And thus, the new hospital should be away from noisy and polluted areas such as industrial areas and dumping sites.
3- Socio-economic factors: these factors mainly includes proximity to transportation and main roads.

D. Assign Criteria Weights (AHP)

In this step, we assign weights and scores for the identified factors in the previous step based on AHP. AHP has been widely exploited in health-care and medical related problems. The following steps are used to assign weights for all identified factor based on AHP:

1- Layout and expose the overall factors.
2- All factors can be compared using pair wise comparisons in order to generate weights for factors through distributing questionnaires on experts. In pair wise comparisons, we decide which factor is more and how much important than another using 1-5 scale as shown in Table 2 [13]. The produced wrights quantify the importance of factors in the analysis and decision making process.
3- Check the consistency ratios of all pair-wise comparisons.
In this step, we use the Consistency Index (CI) and Consistency Ratio (CR) formulas to check the consistency as follows:

$$CI = (\lambda_{max} - n)/(n-1) \qquad (1)$$

Where:
n: the number of criterion.
λ_{max}: the biggest eigenvalue of the comparison matrix.

$$CR = CI/RI \qquad (2)$$

Where:
RI: a constant corresponding to the mean random consistency index value based on n.

4- The relative scores are aggregated using geometric mean method.

Table 2: Pair wise comparison scale

Verbal judgment	Explanation	Number
Extremely Un-Important	A criterion is strongly inferior to another	1/5
Moderately Un-Important	A criterion is slightly inferior to another	1/3
Equally Important	Two factor contribute equally	1
Moderately Important	Judgment slightly favor one criterion over another	3
Extremely Important	Judgment strongly favor one criterion over another	5

E. Data Analysis

In this step, a model is developed in order to identify the optimal location for building a new hospital. In this model, the raw data should have the same spatial reference and they are converted into a raster with the same cell size, making them easier reclassified in Analysis steps.
The data anlysis steps and tools are detailed as follows.

- Distance to existing hospitals based on network analysis: as detailed earlier, there are six factors taken into account for building our model. In this step, we derive a series of polygons (service areas) that represent the distance that is required to reach each hospital. As a prerequisite to finding the service areas, we need to construct a network dataset. The results of applying this step are depicted in Figures 4 and 5.

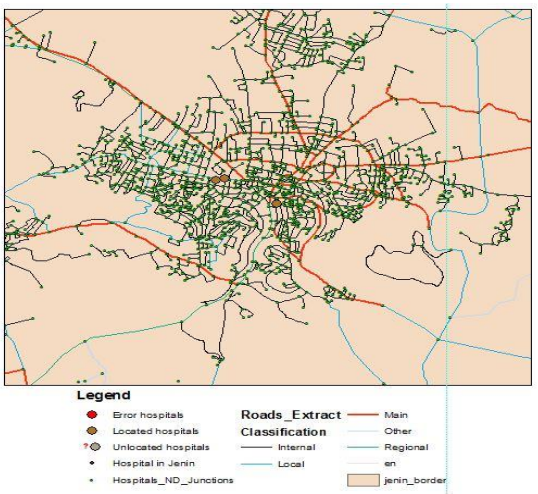


Fig.4: The result of creating a Network Dataset for Jenin roads and facilities.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

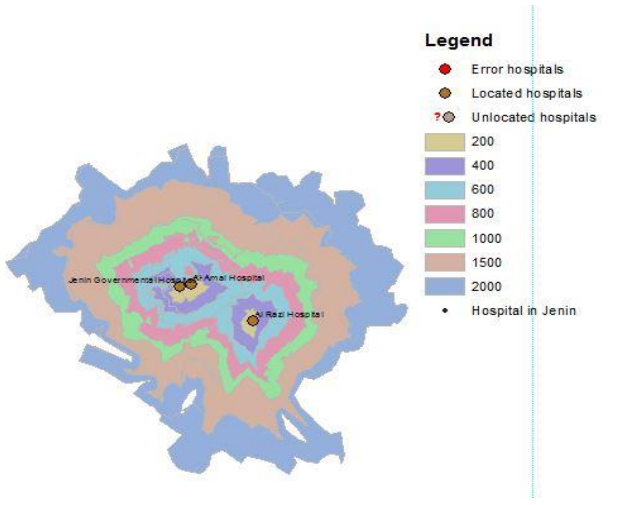


Fig.5: Distance to hospitals based on Service Area polygons.

- Euclidean distance: In this step, we derive the Euclidian distance from facilities (dumping sites and industrial areas) to each pixel in the generated output raster. The formula for finding the Euclidian distance is depicted below:

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(qi - pi)^2} \quad (3)$$

Where:
p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space.
d: Distance from p to q
The results of applying this step are depicted in Figures 6 and 7.

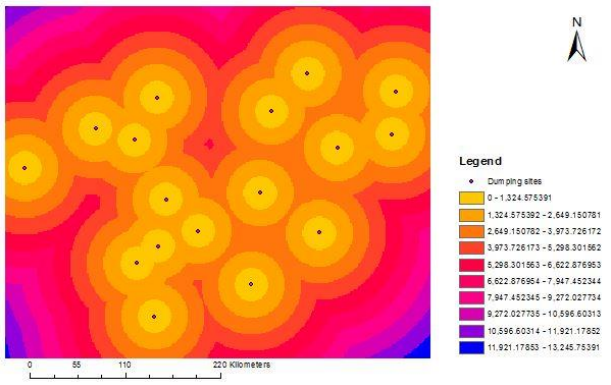


Fig.6: Euclidean distance to dumping sites

- Feature to raster: During this step, we convert the land use feature class (vector data) to a raster that has the same cell size as the derived raster layers from the previous step. Accordingly, we can use all of them for further processing.

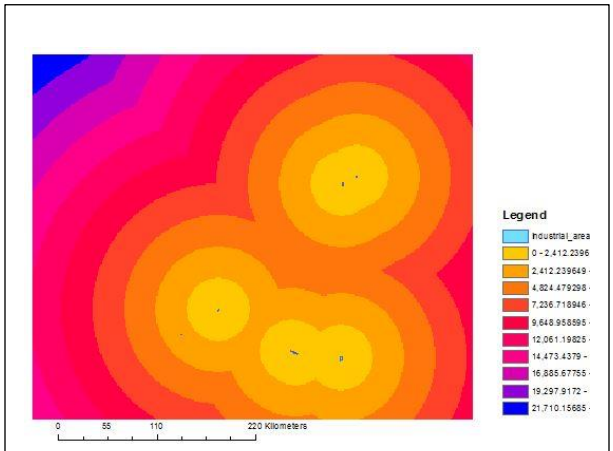


Fig.7: Euclidean distance to industrial areas

- Slope: In order to build a hospital, the land should be relatively flat. Therefore, we consider the slope of the land in our model by deriving the slope of the elevation dataset as shown in Figure 8. By this, we mean that the rate of maximum change in elevations is calculated.

- Reclassification: Each cell in the study area now has a value for the following factors (existing hospitals, dumping sites, industrial areas, land use and elevation). We should combine the derived datasets in order to identify the potential location for building a new hospital in the next step (Weighted overlay). However, we cannot combine them in their current form. For example, there is no meaning to combine cell values that have 15 degrees slope with cell values that have agriculture land use that equals (6). Accordingly, to combine datasets, we need to derive a common measurement scale such as from 1 to 10. This scale identifies how suitable a specific location for building a new hospital. Lower values indicates

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

locations that are more suitable. The reclassification process is depicted in Figure 9.

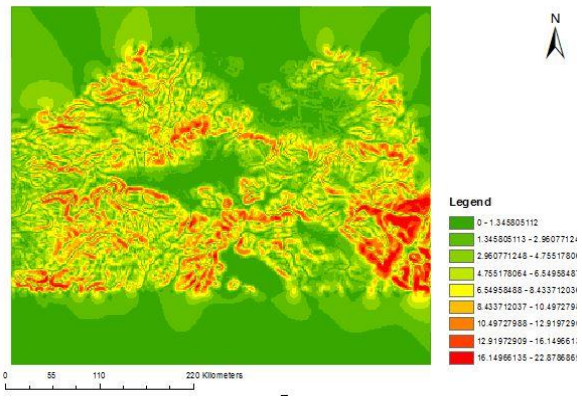


Fig.8: Slope of Jenin governorate

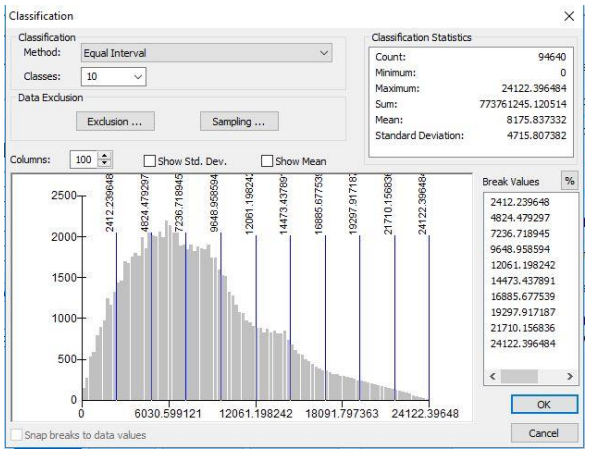


Fig.9: Classification of Euclidean distance values to classes from 1 to 10

As shown in figure 9, we classify the produced distance values from the Euclidean distance process into 10 classes by dividing the produced ranges into equal intervals.

- Weighted overlay: Using this technique, we weight the values of each dataset by assigning each a percentage of influence. The higher the percentage, the more influence an input has in the suitability model. Some input values will be restricted. For example, areas that belong to "C" administrative division in West Bank are restricted as shown in Figure 10. The result of this step indicates how suitable each location for building a new hospital.

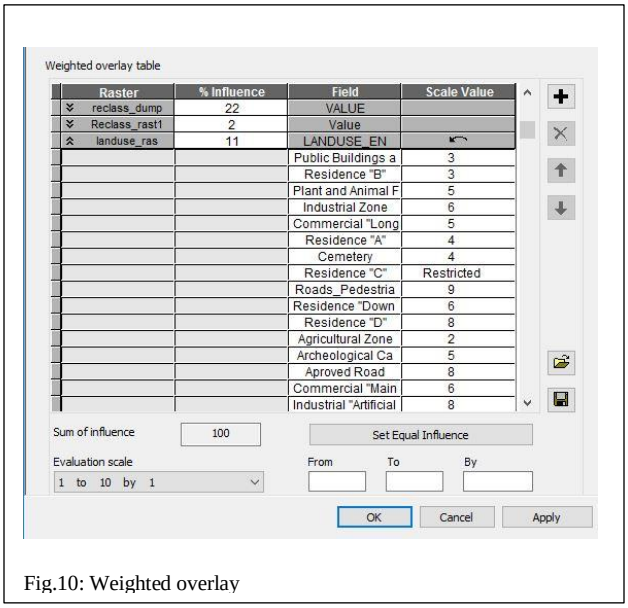


Fig.10: Weighted overlay

- Majority filter: The size of the suitable area is an important criterion in identifying the optimal site for building a new hospital. Thus, we use this tool to ensure that the number of neighboring cells of a similar value must be large enough to build a new hospital.

- Condition: Pixels with values less than 3 indicates suitable locations. Thus, the condition tool used to identify those locations.

V. Experimental Results and Discussion

This section describes the experiments carried out to evaluate our proposed suitability model. A prototype of the proposed model is implemented and experiments are conducted using a PC with dual-core CPU (2.1GHz) and (8 GB) RAM. The used operating system is Windows 10.

- Comparing the Produced Results When Changing the Weights Assigned for all Input Data Sets

In this section, we compare between the produced results by the proposed model when we change the weights assigned for the input data sets according to the following table.

Table 3: The weights assigned for input datasets in different experiments.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Input data set	Random weights in EXP.1	Random weights in EXP.2	AHP weights in EXP.3
Land use.	0.09	0.10	0.22
Distance to existing hospitals	0.04	0.20	0.13
Near main roads	0.24	0.04	0.30
Distance to dumping sites	0.13	0.06	0.22
Distance industrial areas	0.30	0.30	0.02
Elevation	0.20	0.30	0.11

As shown in table 3, we assigned different weights for each input data set in different experiments. After running the proposed model with weights from the first experiment (EXP.1) and the second experiment (EXP.2), we get the results shown in Figure 12.

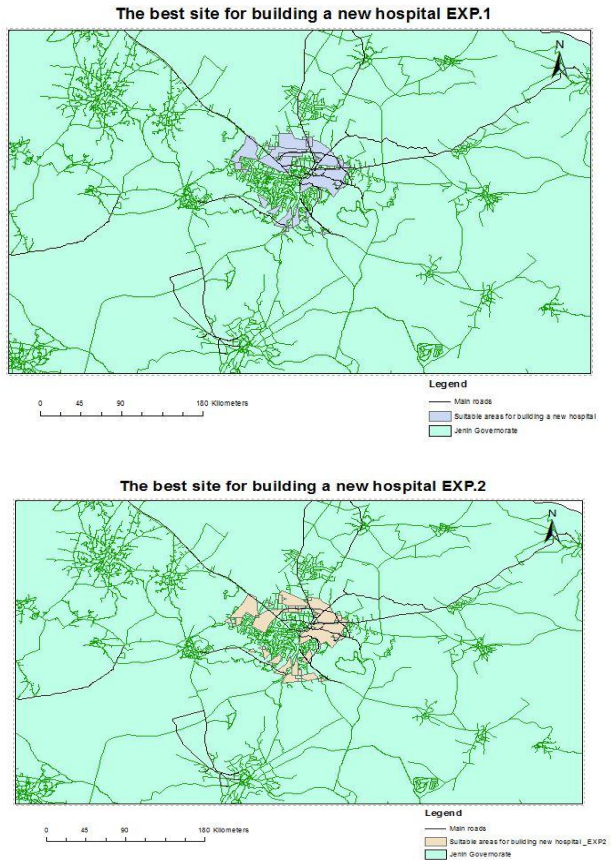


Fig.12: The optimal sites for building a new hospital with weights from EXP.1 and EXP.2 respectively.

As we can see in Figure 12, the produced results have various contiguous alternatives, which may be confusing for the decision maker. This is because the weights are randomly assigned for the input datasets without extensive care or research.

However, we were able to achieve promising results for decision makers when using the weights produced from applying AHP as shown in Figure 13.

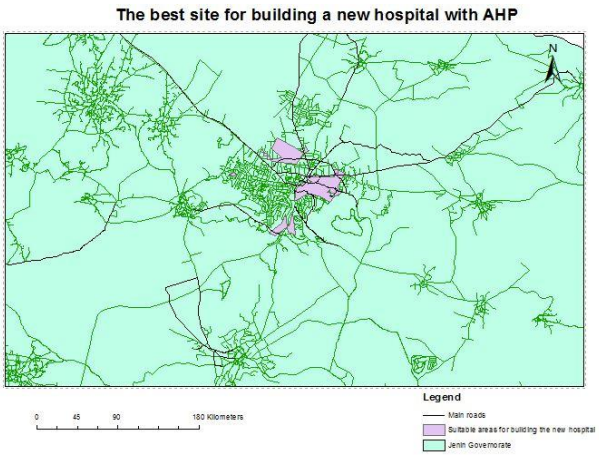


Fig.13: The optimal sites for building a new hospital with AHP weights.

VI. Conclusion and future work

In this paper, we built a suitability model for selecting the optimal site for building a new hospital based on coupling GIS-based MCA and AHP. GIS tool and techniques are employed to analyze the list of identified criteria in hospital site selection. The analysis process incorporates assigning weights for the identified criteria based on AHP. And at the end of the analysis process, the optimal site for building a new hospital is identified. The results showed that assigning weights based on AHP is better than assigning weights randomly for the set of identified criteria.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

References

[1] Liu, Y., 1998. Visualization the urban development of Sydney (1971–1996) in GIS. In: Proceedings of 10th colloquium of the spatial information research centre, University of Otago, New Zealand, 16–19 November.

[2] Lina Zhou, Jie Wu. GIS-Based Multi-Criteria Analysis for Hospital Site Selection in Haidian District of Beijing. 2012

[3] Hare, T.S., Barcus, H.R., 2007. Geographical accessibility and Kentucky's heartrelated hospital services. Applied Geography 27, 181–205.

[4] Astell-Burt, T., Flowerdew, R., Boyle, P.J., Dillon, J.F., 2011. Does geographic access to primary healthcare influence the detection of hepatitis C? Social Science & Medicine 72, 1472–1481.

[5] Gordon, A., Womersley, J., 1997. The use of mapping in public health and planning health services. Journal of Public Health Medicine 19(2), 139–147.

[6] Maglogiannis, I., Hadjiefthymiades, S., 2007. EmerLoc: location-based services for emergency medical incidents. International Journal of Medical Informatics 76, 747–759.

[7] Nobre, F.F., Trotta, L.T.F., Gomes, L.F.A.M., 1999. Multi-criteria decision making: an approach to setting priorities in health care. Symposium on statistical bases for public health decision making: from exploration to modeling18, 3345–3354.

[8] Ohta, K., Kobashi, G., Takano, S., Kagaya, S., Yamada, H., Minakami, H.,Yamamura, E., 2007. Analysis of the geographical accessibility of neurosurgical emergency hospitals in Sapporo city using GIS and AHP. International Journal of Geographical Information Science 21(6), 687–698.

[9] Cheng-Ru, W., Chin-Tsai, L., Huang-Chu, C., 2007. Optimal selection of location for Taiwanese hospitals to ensure a competitive advantage by using the analytic hierarchy process and sensitivity analysis. Building and Environment 42, 1431–1444.

[10] Vahidnia MH, Alesheikh AA, Alimohammadi A. Hospital site selection using fuzzy AHP and its derivatives. Environ Manage. 2009 Jul;90(10):3048-56.

[11] Palestinian Central Bureau of statistics.

[12] http://www.pcbs.gov.ps/Portals/_Rainbow/Documents/jenna.htm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Big and Connected Data Analysis with Graph and Relational Databases Using Collaborative Filtering Technique

Joseph George Davis
Department of Computer Science
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana
spomegasys@gmail.com

Joseph K. Panford
Senior Lecturer, Department of Computer Science
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana
jpanford@yahoo.com

James B. Hayfron -Acquah
Senior Lecturer, Department of Computer Science
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana
jbha@yahoo.com

Abstract- Over the years systems or applications in existence have been able to work seamlessly with relational databases. Applications such as point of sale, hospital management systems, Human resource applications, payroll systems, banking with relational databases banking systems just to mention a few are powered by relational databases with minimal to entirely no issues because most the relationships amongst entities in these systems are mostly not complicated or highly connected. In the same vein relational database systems have been able to handle large amounts of data and transactions emanating from everyday operations of these systems. In summary the relational database was very effective in dealing with the problems applications available were solving. In these modern times the influx of social media platforms, map and navigations systems, geospatial information systems, recommendation engines ,referral systems and the likes have turned the tide for systems and the databases behind them to support, manage and model mostly semi/unstructured, connected and their complex associations amongst data elements. This study looked at such data with the concept of recommendations to test whether relational databases were still performing well with the current trends on connected data or the NoSQL paradigm making inroads in the technology space had a point when they say "they are the database paradigm for the future". An experiment was therefore performed with a relational and graph database to ascertain this analogy.

Key performance indices (KPIs) such as runtime, storage, schema flexibility, query complexity and general operations of each database paradigm were tested with the concept of making recommendations to a number of people in the database based on how their associated(friends) with each other. Each paradigm was put on similar test against the above KPI's and the graph database seemed to have an urge over the relational database in the result and analysis of the figures that were obtained from the experiment.

Keywords: Key Performance Index, Graph Databases, NoSQL, Relational Databases, DBMS

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

I. INTRODUCTION

The battle between relational [9][11] and non-relational databases began a while ago but technological and software application trends happening now and also what is perceived for the future has intensified this battle. Applications now are thriving more on information which relates to how stored data is connected and associated with each other for further analysis, decision making and recommendations. A direct example is the social media platforms available nowadays hence database management systems behind these systems have to be in the position to meet these ever changing trends and requirements [6]. Each database paradigm (relation and non-relational) has dealt with this issue in its own peculiar way which may or may not have worked in certain circumstances. So the question still remains which of these database systems better handles connected, associated data better for applications that need to interpret this data for analysis and decision making to work in a fast and efficient way, using less processing power, memory, storage space and with less complex queries[1]. The problem most often sterns from when the connected and associated data being used by these applications (social media platforms, map and navigation systems, geospatial information systems, recommendation engines (mostly behind e-commerce platforms), referral systems, Internet of things starts growing and complex questions and analysis are requested from the applications to drive short term to long term strategic decision making for organizations. . It is in such situations that the ACID [9] (atomicity, consistency, isolation and durability) features of a database are really tested. So the question still remains which type of database systems (relational or non-relational) is for the future in terms big and connected data which seems to be the bane of applications coming up each single day.

This study therefore aims to test how each of the database paradigms will behave in terms of the parameters noted below.

- General operational differences
- Storage analysis on both DBMS
- Runtime analysis of both DBMS interacting with associative data
- Query complexity analysis
- Handling of schemas for connected data sets.

II. JUSTIFICATION OF THE STUDY

The purpose of this study is to test and profile a relational [9][11] and graph database[16] working behind a recommender application which works on analyzing the associations between data constituents for decision making and strategic planning. In other words, scenarios will be setup where by people having relations with others in the database are suggested products that friends have bought or liked using relational and graph queries[1][11]. In a nutshell future applications which will be depending mostly on strong analysis of connected data will be built knowing which database paradigm will support and answer the complex questions that will be posed to such applications.

On a higher level application throughput will be tested as query time for complex questions/situations which are solved by such applications tracing and analyzing patterns in user associations in data presented to predict future behavior or preferences. Also amount of physical storage to be used by DBMS(relational and graph database) as data grows in certain proportions in such situations will also be a question to be answered as well as how fast the application will run when data increases. In other words which database paradigm will survive with large amount of data with complex associations? Moreover, a test will also be made if the combination of the two databases backing the same application may or may not help solve complex situations better. This will base on leveraging the strengths of each database uncovered in the course of the study to mitigate the shortfalls of the each other.

III. LITERATURE REVIEW

Thought relational database concept proposed by E.F. Codd [9] has thrived over years and has stood the test of time giving it a merited status as a matured database system, [6] with the recent introduction of systems like the World Wide Web
[15], social media, social networking applications [7], internet of things and the likes have put this maturity to test, raising questions of in database domain in relation to flexibility when it comes to data sets used in these modern applications[8] . This trend has caused the emergence of new database paradigms like NoSQL [14] trying to mitigate the short falls of the relational databases in this regard has a lot variants of databases from columns store to key-value database systems. This study looks at the graph database in relation to the relational database on how they handle related data objects for systems that need it for decision making. Related studies like one done by Emil Eifrem [4][12] looked at comparing graph and relational database in regards to a social example of a 1000 customers/users with an average of 50 friend connections with degrees of connections giving friend-of-friend, friend-of-friend-of-friend down to the 4th hop scenarios. Queries run against each database produced execution time of 2ms for graph and 2000ms for the relational database for a small dataset. The scenario got worse for the relational database since it had to be stopped after a day of running the same query which the graph database run in 2ms.From the study the graph database proved better in that regard.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Looking at a similar research by Sharma M. & Soni P. [1] to analyze the performance of both relational and graph databases with some predefined data processing queries based on a schema to analyze data processing and analyze time with various number of inputs of data objects. The schema used was modeled around a user, friend, movie, actor entities with the queries that need to get all friends of a user, favorite movies of a user's friends and finding lead actors of the movies watched by a user's friends. The study goes on to run these three queries against increasing datasets of 100,200,300 with the relational database having a greater runtime than the graph database at each three instances. B. Shalini & Tyagi C [6] evaluates graph (Neo4j) and relational (MySQL) databases based on evaluating parameters like level of support/maturity, security of the database and database flexibility. In terms of maturity an argument is made for the relational database for providing storage and robust support for commercial applications or products for over a decade as compared to graph databases which came into the technology limelight in the early 2000s so may not be at the same wavelength as relational databases when it comes to production testing over a long period of time. Looking at the security view point the relational database with its built-in multiuser support feature and restrictions is in better shape than the graph database because of its comprehensive support for Access Control List (ACL). Nevertheless the graph database takes lead when it comes to flexibility because of the short fall of relational databases to extend schemas or databases [10].Relational database also lags when it comes to management of flexible schemas that change over time [13].As most studies seem to always target just the runtime of queries, this study also looks at that and extends the parameters to add storage analysis, compare schemas and investigate query complexities

IV. METHODOLOGY

For the analysis of connected data one has to know what connected data is. It is mainly data that has individual entities interconnected with each other such that decisions and analysis are made based on the connected relationships between the entities[6][12]. The focus of this study is on such connected data and how it is analyzed by graph and relational databases taking into considerations the size of the data or rows involved from thousands of records, all the way up to a million interconnected records.

This study will be centered around a social relationship [7] amongst friends based on which decisions or recommendations will be made to highlight the base concept of connected data. The recommendation technique used in this study to inform the composition of graph and relational database queries is the Collaborative Filtering (CF). CF also referred to as social filtering, filters information using recommendations of other people (mostly friends and acquaintances). For example a person who wants to watch movie may ask for recommendations from friends. The recommendations of friends who have similar interests are trusted more than recommendations from others [18].The data used for the study have been modelled around CF concept to ascertain which of the database paradigms works well in situations where connections among data entities are used to make recommendations .The data that will be looked into can be from a wide variety of products from a hypothetical ecommerce site but it will be towards users purchasing movies, books, games and electronic gadgets. Figure 1 is the underlying general data model.

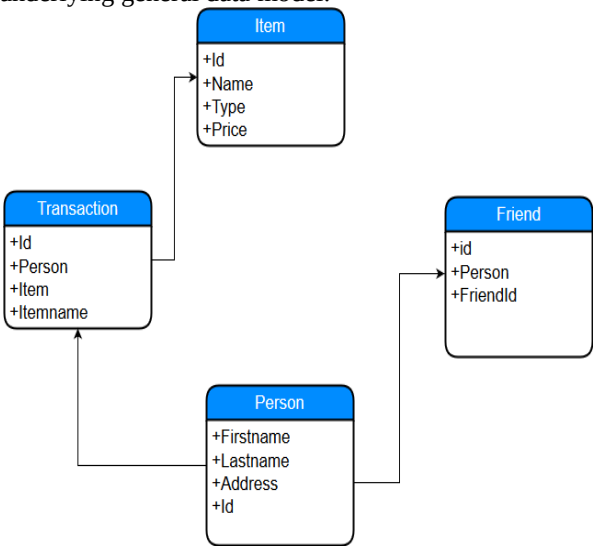


Figure 1: The general data model irrespective the type of database used.

Figure 1 is the underlying general data model which projects four entities (Person, Friend, Item, and Transaction) for the study. This data model portrays people and their friends, how they relate to the items in a hypothetical online store from which recommendations will be made by the system using CF. A person buys or likes an item (s) be it a movie, book, game or electronic gadget. A person also has friend(s) who also has liked or bought an item(s).People can have a lot of friends, buy or like a lot of items so we have large amounts of data with entities associating or relating to each other .The question is how do we leverage on the features of a relational or graph database to find patterns in data that can help recommend or suggest items and even friends to people.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

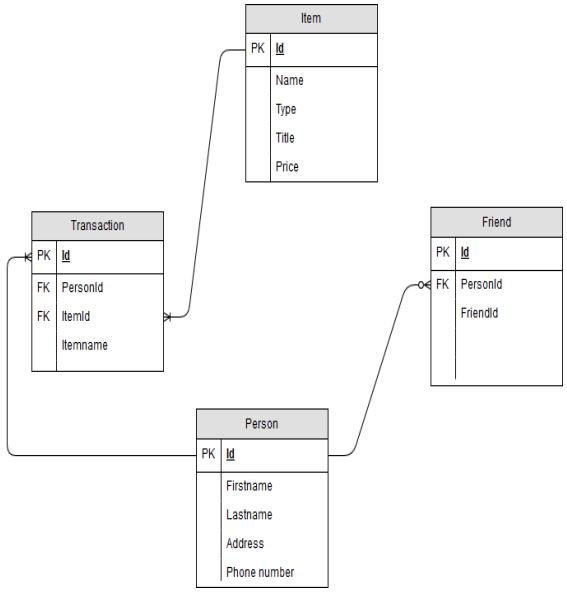


Figure 2: Relational Database Model

Figure 2 shows the relational data model for the system. This shows the relationships the person entity has with the other entities through the foreign key constraint. This further elaborates that a person has friends and can buy or like an item which can be a book, game, movie or a gadget etc. In the relational database paradigm [9] this relationship is shown with a foreign key linking the person entity to the other entities. This is how relationships are portrayed in the relational sense showing connectivity amongst individual entities.

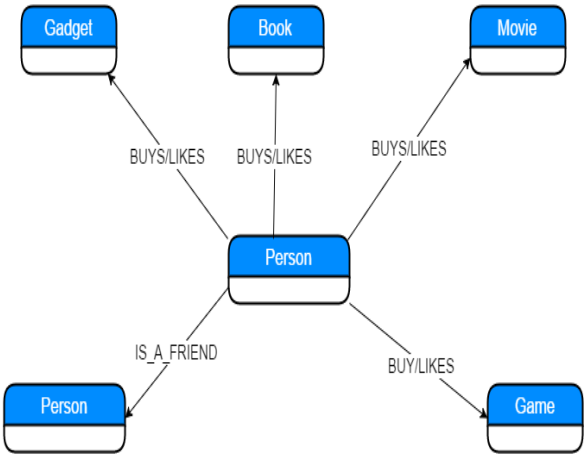


Figure 3: Graph Database Model

The figure 3 above depicts how data modeled in graph [2][5][10] looks like. In graph sense each entity is represented by a node with its properties being the attributes of that node. The relationships between nodes are represented by the labeled vertices between them [16].
This study looks at how these relationships between people, their friends and products can be leveraged to answer questions like "What are the products a person's friends are buying" or "What products are a person's friend of friends buying" which is the basis of recommendation systems to recommend products to people based of their relationship and buying patterns of friends, friends of friends etc.
The system implementation in code was based on the below architecture. The system comprises of a JAVA Enterprise Edition Application riding on a graph (Neo4j) and relational (MySQL).

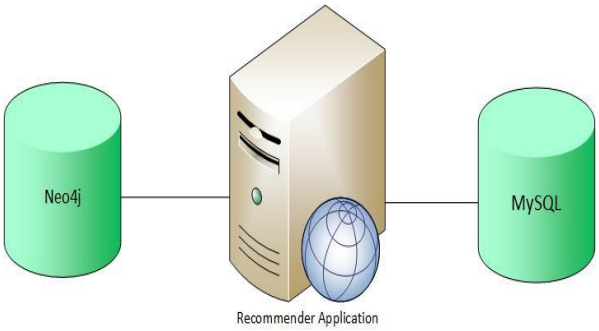


Figure 4: General System Architecture

From figure 4 the system comprises of:
I. Application Server is a JAVAEE application deployed and running on a Glassfish server where the business logic for the application resides. This is where all decisions are made using the information obtained from the attached databases
(MySQL and Neo4j).
II. Neo4j [16][13] Graph Database is Graph DBMS that interfaces with the graph implementation side of the Application.Neo4j can be implemented in server mode or embedded mode. For this study/application it is used in embedded mode meaning the management of the database is embedded in code using the Neo4j libraries this means where ever the code is deployed an instance of the Neo4j database is created.
III. MySQL is the Relational DBMS that interfaces with the relational implementation side of the Application.
The systems also works with the below functional requirements:
I. This system works with a graph and relational database to help make product suggestions to people based on the pattern of purchases or likes other people and friends have been doing (Collaborative Filtering)[14].
II. Based on the behavior in (I) above each database will be populated and tested with increasing amounts of data (i.e10000 people,100000 people or 1000000 people) to answer questions like "what have friends of a person bought or liked so that recommendations can be made for that person.
III. Analysis on the runtime, amount of storage used and complexity of queries from the database systems used from the data and questions posed to the system in (II) above will be looked at.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Any system that the application can be deployed on must satisfy the requirements below
Software:

- JAVA JDK 7 and above
- MySQL RDBMS
- Neo4j is run inembedded mode
- GlassFish version 4 and above

Hardware: System will be tested on a computer with the specification below

- Intel Core i3 processor
- 8GB RAM
- 500GB hard disk space
- OS Linux Ubuntu 16.04

V. RESULTS

Looking at the runtime analysis of the same query run with a start of hundred transactions all the way to a million with the query average runtime calculated in each scenario The below results was attained for relational & graph databases:

Table 1: showing the runtime against data results for relational database

data	Execution time/secs
100	0.0028
1000	0.016
10000	0.164
100000	1.315
500000	6.244
1000000	12.896

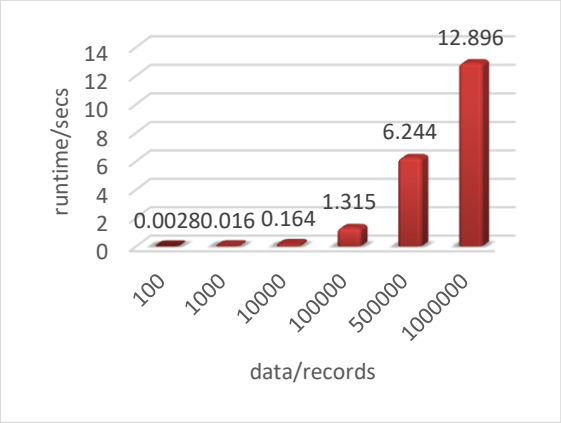


Figure 5. Chart showing amount of data run in query against runtime for a relation database

1

1

Table 3: showing the storage against data results for relational database

data	Storage space/kbs
100	792
1000	840
10000	2744

Table 2: showing runtime against data results for graph

data	Execution time/secs
100	0.002
1000	0.006
10000	0.09
100000	0.219
500000	1.616
1000000	1.728

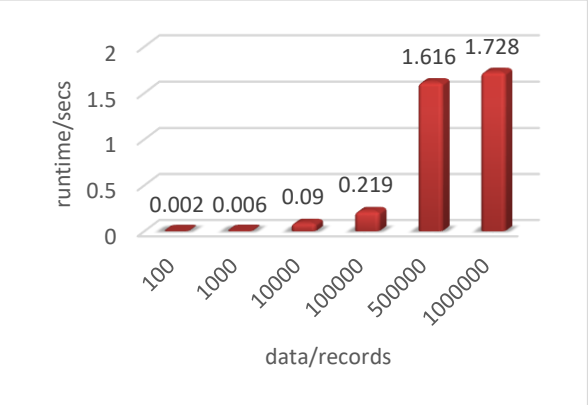


Figure 6: Chart showing amount of data run in query against runtime for a graph database

Figure 7 shows runtime analysis trend of relational and graph databases as data increases.

Figure 7: Runtime analysis trend

The same analysis is done with the storage space used by each database shown in figure 10.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

100000	14024
500000	37596
1000000	58096

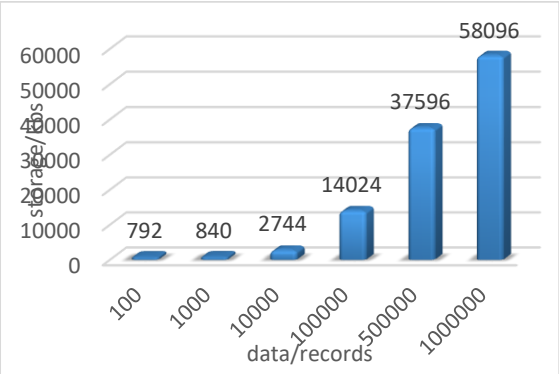


Figure 8.Chart showing database storage size against record size all in kilobytes

Table 4: showing results for graph

data	storage/kbs
100	400
1000	2000
10000	18200
100000	105100
500000	535000
1000000	839000

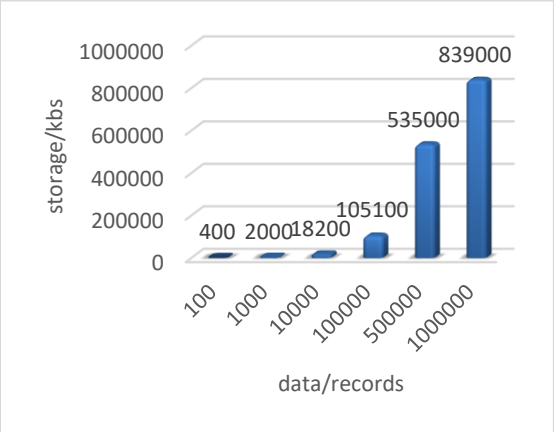


Figure 9: Chart showing database storage size against record size all in kilobytes

Combined Storage Analysis Figure 10 shows the storage analysis trend of relational and graph databases as data increases.

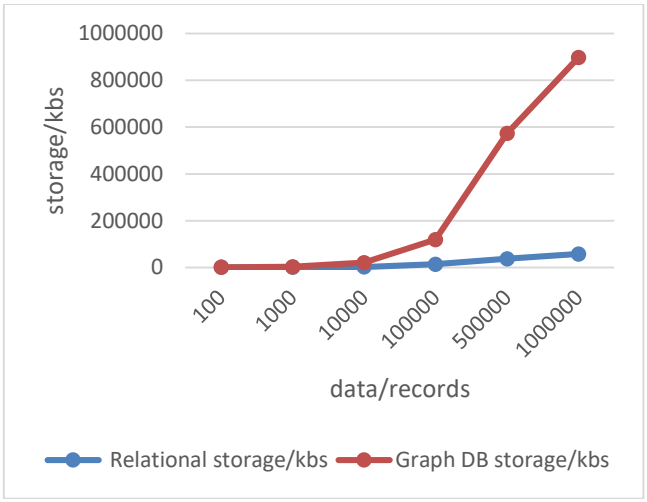


Figure 10: Storage space used by Graph & Relation database

Looking at the trends in figures 7 and 9 it can be seen from the runtime metric trends that as data or number of records increase runtime on each database systems behave differently, for the relational database runtime starts increasing exponentially as the data size reaches 100000 and over hitting as much as 12secs with a million records which is worrying in terms of system performance with large datasets. The graph databases on the other hand maintains a steady increase in runtime as the data sizes reaches 100000 and maintains a runtime of under 2secs even with a million records making it a good option for systems that do large and connected data sets analysis. The story is different when it comes to the amount of space needed by each database to handle large sets of data. It can be seen that the graph database uses a large amount of space as the dataset gets larger whiles the relational database uses less that 60MB to handle a million records. Even though this development is alarming on the part of the graph database, storage does not seem to be problem for current systems and applications because most of the servers or machines come with not less than 500GB.That notwithstanding it is a big plus for the relational database for its storage optimization.When we assess the complexity of the queries on both sides we get to know why the relational
database has a high runtime. From the query complexity analysis it is seen that when it comes to analyzing connected data the relational database often uses joins to associate entities. This presents a complexity of O (n2) because join queries are executed programmatically in nested loops whiles graphs queries have a complexity of O (log n) due to the traversal of the graph data structure. This explains the runtime patterns in figure 7.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI.CONCLUSION

The analysis of the study looked at the runtime of queries for the two database paradigms followed by analyzing the storage space used as the size of grows. Also, how both relational and graph databases model connected data in the quest to create a database structure that will fulfil the objective at hand.

Based the on the analysis done it can be concluded that when it comes to dealing with big and connected data the graph database has an urge over the relational database. This makes graph database a great options for future applications which will make decisions based on complex data associations and cause to worry on how far the relational paradigm can carry us. Relational databases were conceived to digitize paper forms and automate well-structured business processes, and still have their uses. But RDBMS cannot model or store connected data and its relationships without complexity, which means performance degrades with the increasing number and levels of data relationships and data size. Additionally, new types of data and data relationships require schema redesign that increases time to market.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] Mukul S. & Pradeep S. (2014), Quantitative Analysis and Implementation of Relational and Graph Database Technologies, International Journal of Modern Computer Science and Applications, RES Publications, Vol.2, p: 21-25.

[3] Buerli M. (December 2012), The Current State of Graph Databases, Department of Computer Science Cal Poly San Luis Obispo

[4] Josh A. (2013), Performance of Graph vs. Relational Databases, viewed January 2016 from: https://dzone.com/articles/performance-graph-vs

[5] Angles R. A comparison of current graph database models. In Proceedings of the 2012 IEEE 28th International Conference on Data Engineering Workshops, ICDEW '12, pages 171–177, Washington, DC, USA, 2012. IEEE Computer Society

[6] Batra S. & Tyagi C. Comparative analysis of relational and graph databases. International Journal of Soft Computing

[7] Cudr´e-Mauroux P. & Elnikety S. Graph data management systems for new application domains. In International Conference on Very Large Data Bases (VLDB), 2011.

[8] Vicknair C., Macias M., Zhao Z., Nan X.,Chen Y., &Wilkins D., A Comparison of a Graph

[9] Codd E.F, A relational model of data for large shared data banks June 1970

[10] Angles R. and Gutierrez C. Survey of graph database models. ACM Comput. Surv., 40(1):1:1–1:39, February 2008.

[2]Williams D., Huan . &Wang W, Graph Database Indexing Using Structured Graph Decomposition, Department of Computer Science University of North Carolina, Chapel Hill, Department of Electrical Engineering and Computer Science University of Kansas

[11] Connolly M. T. & Begg E. C.(2005),Database Systems, 4th Edition, Addison Wesley, UK.

[12] Robbinson I, Webber J. & Eifrem E. (2015), Graph Databases, Oreilly, California.

[13] Femy P.F, Reshma K.R,Varghese S.M ,Outcome Analysis using Neo4j Graph Database, International Journal on Cybernetics & Informatics (IJCI) Vol. 5, No. 2, April 2016

[14] Jampana A.K, New Era of NoSQL Databases, International Journal of Emerging Trends in Engineering Research, Vol.3. No.10, [Pages: 362-367(2015)

Special Issue of ICACSSE 2015 -Held on October 30, 2015 in St. Ann's College of Engineering & Technology, Chirala, AP, India

[15] Berners-Lee, Tim. HyperText Transfer Protocol Design Issues, 1991. http://www.w3.org

[16] Neo4j. http://www.neo4j.org/.

[17] Wikipedia. http://www.wikipedia.org/.

[18]Ekstrand M.,Riedl J. & Konstan J, Collaborative Filtering Recommenders Systems Vol.4,No.2(2010),81-173

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3D Facial Features in Neuro Fuzzy Model for Predictive Grading Of Childhood Autism

Reji R[1], Dr P SojanLal [2]

[1]Research Scholar, School of Computer Sciences,

Mahatma Gandhi University, Kottayam, Kerala, India.
reji.r@saintgits.org

[2] Principal, Mar-Baselious Institute of Technology & Science,

Kothamangalam, Kerala, India.
sojanlal@gmail.com

Abstract: Autism Spectrum Disorder (ASD) is a clinically heterogeneous neurological developmental disorder. It is called a spectrum disorder because of its range of symptoms. Early diagnosis and proper intervention is required for the effective treatment of autism. Diagnosis is based on the quantitative and qualitative analysis made by the clinician. The expertise of the clinician is so important in the proper diagnosis and classification of autism. This paper proposes an Expert system that act as a support system to the clinician. Major clinical attributes of autism along with facial features are used as input to the expert system. The main highlight is the use of feautures from 3D facial imagery for autism classification. The expert system operates in two modes, diagnosis mode and grading mode. Naïve Bayes classifier is initially used for diagnosis mode where as overall system is implemented using a Neuro-Fuzzy approach. In the diagnosis mode 100% accuracy and in classification mode 98.8% accuracy is obtained.

Keyword: Autism, ASD, 3D Face, Neuro-Fuzzy System, Neural networks, Fuzzy logic, Expert system.

Introduction: Autism spectrum disorder (ASD) is a clinically heterogeneous condition with a wide range of factors. A satisfactory diagnosis measure for ASD is currently unavailable. Autism is a neurological handicap in children, which is usually diagnosed in early child hood. There is lack of definitive biomarkers for autism diagnosis. The diagnosis mostly depends on a range of factors. People with autism show different clinical features and symptoms. There is lot of scope for quantitative research on ASD in developing countries like India. Satisfactory and accurate data for research in autism is unavailable in India. The frequency of ASD diagnosis is increasing. Many Factors like increased awareness,improved detection mainly contribute to this. The publication in DSM-5 on May 2013 adds major revisions needed to remove the confusing labels associated with ASD. The earliest symptom is the absence of normal behavior. All children should be screened using a standardized Autism screening tool at 18 and 24 months of age [1].

Symptoms of ASD must be present in the earlier developmental period mostly by the second year of life (after 12 months). But least severe type of ASD may be diagnosed by 4 to 6 years or later. Intervention should begin as early as possible. In intervention consider the core distinctive features of autism and it should be specific and proof based. More over it should be well structured and appropriate to the developmental need of the child.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Even though there are inter individual difference in the clinical levels of children with ASD, they share some common characteristics like deficit in social interaction/communication and behavioral abnormalities.

Studies shows that children who deficit to recognize face in childhood shows severe autistic features at teenage. Researches shows that human recognize a person by their body if someone is coming from far away or the face is obscured. So for identifying a person, brain uses facial characteristics and also other physical cues.

Researchers at the University of Missouri have identified facial features measurements in children with autism and developed a screening tool for young children. The sample consists of children from 8 to 12 years of age. Judith Miles, Professor Emeritus of child health-genetics in the MU Thompson Center for Autism and Neuro Developmental Disorders point out that a portion of those children diagnosed with autism tend to look alike with similar facial characteristics [2].

In this research we are developing an expert system that use core clinical features with its attributes, facial characteristics and parental status as input.

Autism: Clinical features and Diagnosis

Autism detection can be done by using quantitative tests and qualitative analysis. In DSM IV ASD diagnosis is based on Language delays, Social Communication Problem and Repetitive behavior. Where as in DSM V ASD diagnosis is having two criteria domains namely Social interaction domain and Repetitive or restricted behavior domain. The Core Clinical features of autism can be brought under the following heads with attributes.

1. Behavioral problem
   a) Poor eye contact
   b) Lack of responsiveness to others
   c) Difficulty in building social relationship
   d) Repetitive acts
   e) Self harm
   f) Compulsive behavior
   g) Hyper Activity
   h) Poor joint attention
   i) Solo play
   j) Excessive fear
   k) Poor emotional response

2. Language Disorder
   a) Muteness
   b) Echolalia
   c) Sound making

3. Intellectual retardation
   a) General intellectual retardation
   b) Brain Seizures

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

4. Facial Features
   a) Open Eyes
   b) Wide Mouth
   c) Large region between mouth and nose.
   d) Expression less face
   e) Open mouthed Appearance
   f) Prominent Forehead

5. Parents Status
   a) Not Autistic
   b) Autistic

The earliest symptom is the absence of normal behavior. Normally when a parent or a healthcare provider notices any delay or abnormal behavior in the child at, or prior to the age of three they are prompted to consult a developmental pediatrician. The child is analyzed carefully and any abnormality is observed in the core functional areas, the developmental pediatrician recommends the child for assessment test using any of the standard autism testing tools. These tools are normally a checklist or questionnaire containing autism features. The clinician fills the data using his observation and a structured discussion with the parent of the child under scrutiny. After filling the details a final score is generated. Comparing the obtained score with the threshold value, the clinician initially classifies the child as either not autistic or autistic. The next step is to identify which Autistic class or grade the child belongs to. Based on the total score compared against a threshold the child is diagnosed as mild, moderate and severe. Consider the total score(S) adds up to 60 and the threshold is 30, the grade and remarks is as shown in table 1.

Score	Class/ Grade	Remarks
Score <30	Normal	Typical
Score 30 to 34	Mild	Requiring support
Score 34 to 38	Moderate	Requiring Substantial support
Score >38	Severe	Requiring very substantial support

Table 1: Score with Grade

The expertise and dedication of the clinician is an important factor while analyzing the grade or class of autism. Expert clinician can easily spot the grade of autism. Some clinician fully depends on the diagnosis tool and there are possibilities of wrong classification. More over the fuzziness in the Score may also lead to misclassification. Studies say that a proper initial diagnosis and follow up is required for autism. If we are using an expert system as a support system for clinicians the misclassification and problems in initial diagnosis of autism can be avoided up to an extent. In this research we are developing an expert system to assist clinicians in their diagnosis procedure.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Related work:

Silberberg et al.[3] focus on the prevalence of neuro-developmental disorder among children aged 2 to 9 years in the different areas of India. They also analyzed the risk factors associated with neuro-developmental disorders along with the development of screening and diagnosing methodology.

An investigation related to the epidemiology of ASD in India was reported by Mukerji et al.[4]

Myers et al.[5] suggests that the primary goal of treatment for ASD is to maximize the child's ultimate functional independence and quality of life by minimizing the core features of ASD.

Robins et al.[6] objective is to validate the modified checklist for Autism in toddlers.

Yasmin H. Nuggers[7] studied the prevalence, risk factors and diagnosis of ASD in developing countries. In his brief reviews controversies regarding the increase in estimate of prevalence, implications of changes in ASD definitions are also discussed.

Vijay Sagar KJ[8] focus on the study of developmental disorders in India. He concludes his article by saying that there is a need of proper diagnosis and screening tools for Autism in India.

Hammond et al.[9] proposes the use of dense face models in 3D Analysis of facial morphology. The model provide a detailed visualization of 3D face shape variation with capability to training the Physicians to recognize the core components of particular syndromes. Ten fold cross validation testing is done on the sample faces using different pattern recognition algorithm.

Vezzetti et al.[10] highlights 3D human face descriptions, land marks measures and geometrical features. Analysis of facial morphology is very important in the study of facial abnormalities.

Gupta et al.[11] worked on the assumption that different facial expressions can be considered as isometric deformation of facial surfaces .Even though deformation occurs, the intrinsic property of the surface remain the same.

Aldridge et al.[12] investigation focus mainly on the correlation between brain development and face. Brain develops in concert and coordination between the developing facial tissues. ASD is due to alteration in embryological brain, suggests that there are differences in the facial structures of ASD children and normally developing one. Finally the authors concludes that there are significant differences in the facial morphology of boys with a ASD compared normally developing one.

Weigelt et al.[13] reports the face identity recognition is deficit in ASD. The deficit is both process specific and domain specific. They suggest that Autism is a domain specific disorder.

Ruggeri et al.[14] objectives is to find the similarity and difference between the terms biomarker and endophenotype. There study includes the established biomarkers and endophenotype in autism research along with the discovery of new biomarkers.

Dataset: The background study and data collection for this work is done at Block Resource Centre Cherthala, Kerala, India. BRC is a Government agency working along with Sarva Shiksha Abhiyan. The dataset consists of 47 children, which includes both boys and girls. The ratio of boys and girls is 12: 1. The age is from 2 years to 12 years. While studying and analyzing the dataset we are making use of the expert opinion from Pediatric Neurologist, Developmental Pediatricians, Speech Therapist, Remedial Educators, Clinical Psychologist and Parents.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Proposed system with objectives:

Objective: Our research focus on developing an expert system for the initial diagnosis and grading of childhood autism. This system can be used as a support system for the clinicians while diagnosing autism. The proposed system is having two modes of operation, Diagnosis mode and Grading mode as shown in figure 1. Initially in the diagnosis mode expert system predicts whether the child is non-autistic or autistic. Once the output of the diagnosis mode is autistic then the next phase is activated. In this phase a detailed analysis is done and the possible outcome is the class or grade of autism.

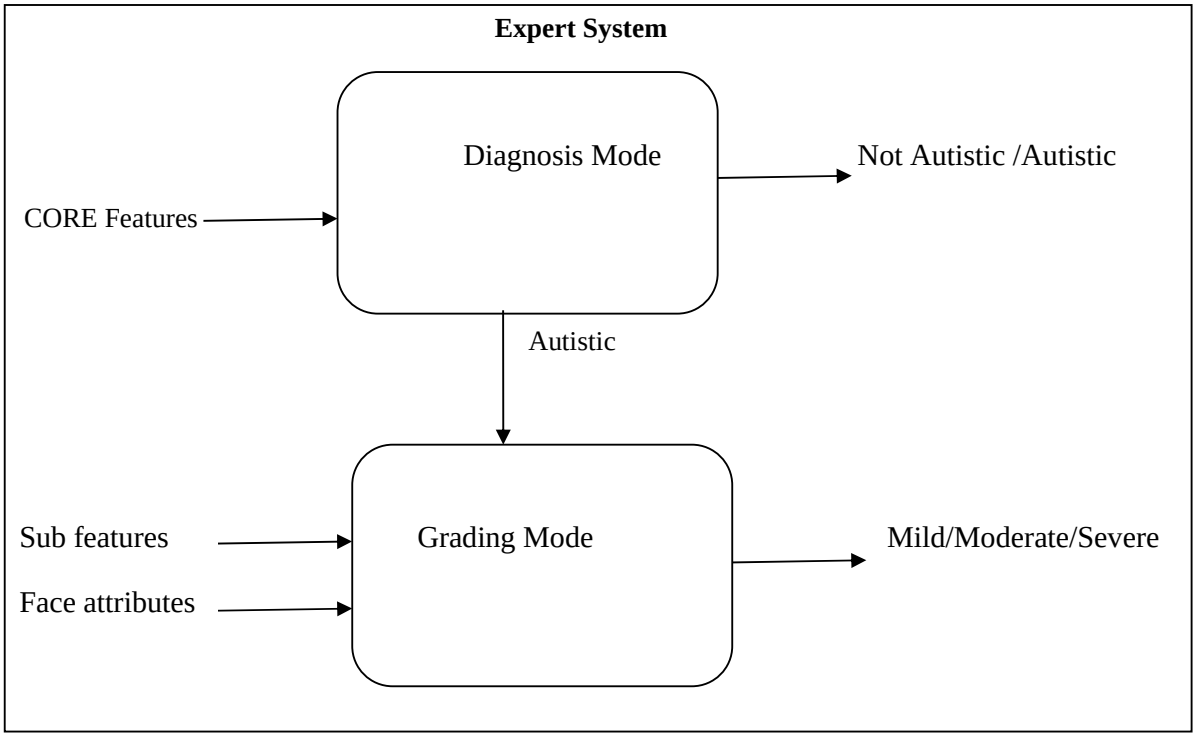


Figure 1: Flow chart of the expert system.

Scale	Output	Remarks
0	Normal	Non autistic
1	Mild	Requiring support
2	Moderate	Requiring Substantial support
3	Severe	Requiring very substantial support

Table 2: Grading

The core feature of autism is analyzed initially during the diagnosis mode. Core features includes Behavioral problem, Language disorder and General mental retardation. Based on the core features the diagnosis mode output is not autistic or autistic.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

If the output is autistic then the second phase is activated, the sub features of each of the core features are analyzed. Facial features along with status of parents are along given importance. The most important characteristic of this expert system is the integration of facial features. The face image of the child under diagnosis is captured; it is modeled to 3D or captures the image using a 3D imaging system. In 3D imaging the geometric depth information is having more importance. Facial features include mouth, eyes and the region between mouth and nose. A portion of Autistic children's have wider mouth, open eyes and large region between the mouth and nose as shown in figure 4. Other common facial characteristics are expression less face, open mouthed appearance and prominent forehead region. Using 3D Geodesic distance as the measure identifies the variance of the features from normally developing kids. Our aim is to extract the exact geometrical information from the face under scrutiny and compare it with a template and used this information for training. By using these facial attributes our focus is to study the contribution of each feature to grading of autism.

Our analysis, point out the fact that Children below the age of 8 with other clinical features of Autism mostly lack the facial features mentioned above. But children from the age of 8 and 12 have shown the above mentioned facial features along with other clinical features of Autism. Our expert system is designed in such a way that the weightage of facial features is varied by considering the age of the child under diagnosis. If the age is below 8 the weightage of the features in percentage is as 75(core features) : 15(facial features) : 10 (Parents status). Whereas age range from 8 to 12 the weightage of the features in percentage is as 65(core features) : 25(facial features) : 10 (Parents status).Parent status is also considered, this feature include whether the parents are autistic or not and age of the parents during conception is also given weightage.

In the grading phase three sets of features namely attributes from core features, facial attributes and parent's status is considered. The weightage of the features varies depending on the age of the child. Initially we consider the two phases as two separate classification problem. In phase 1 the number of inputs are limited so a Naïve Bayes classifier is applied and it suites our problem and it gives the result autistic or non autistic as shown in figure 2. The input to the classifier is the core features such as Behavior problem, Language Disorder and General Mental retardation.

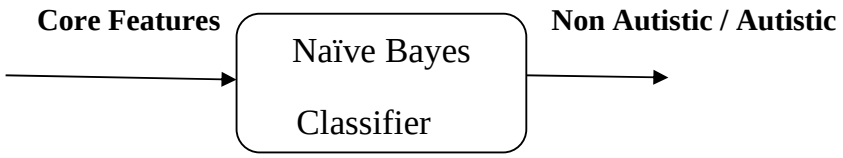


Figure 2: Diagnosis Mode

In the second phase more inputs belonging to different features are considered which include attributes from the core features, facial region and parental status. Naïve Bayes classifier is applied and result is analyzed but there exists some fuzziness after a certain threshold. We need to integrate the two phases and a neuro- fuzzy approach is applied. Soft computing approach like neural network and fuzzy logic can play a vital role in the design of such an expert system. Fuzzy logic is used to interpret expert knowledge directly using rules with linguistic base. In this system we are qualitatively collecting lot of information with structured discussion with parent and from clinician's observation.

Linguistic base can easily be framed into fuzzy rules. Neural network are good in recognizing patterns. So this hybrid approach yields better performance. The output of the grading phase is as shown in table 2.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Results:

To design the neuro-fuzzy system for diagnosis of autism we consider the attributes of core features, facial attributes and parental status. The hybrid architecture is as shown in figure 3.

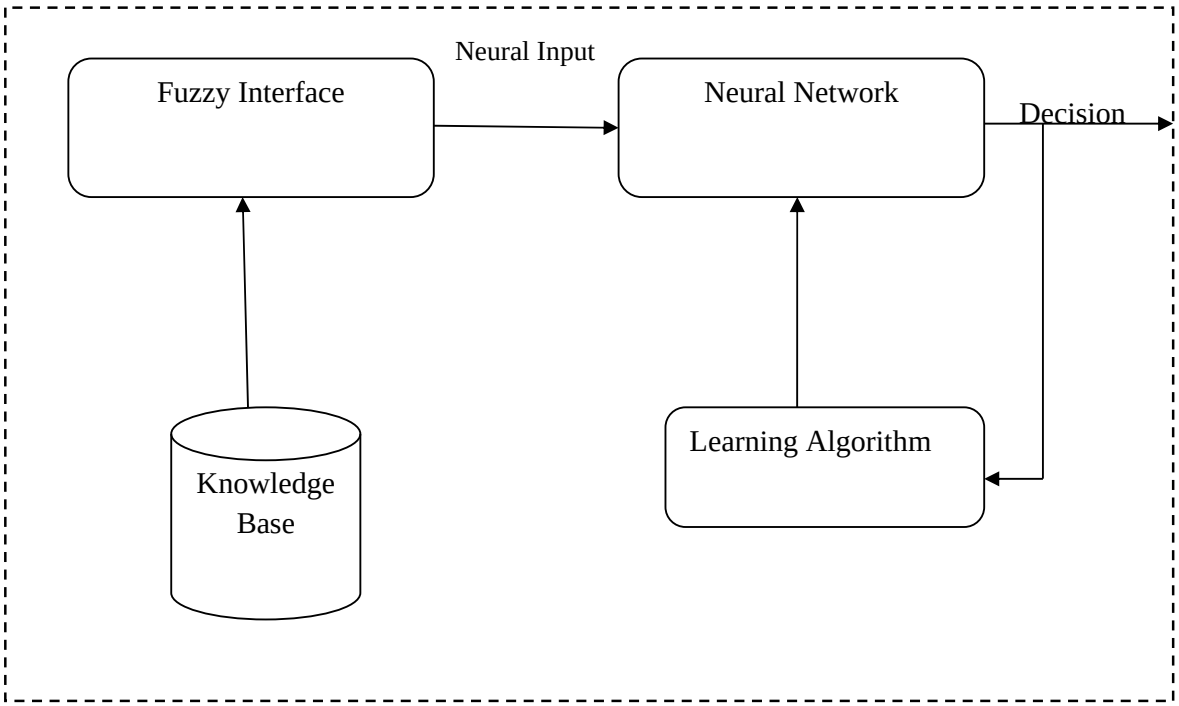


Figure3: Neuro-Fuzzy Expert System Architecture for predicting Autism

The knowledge base consists of twenty two fuzzy parameters. The neural network is trained to learn the parameters of the membership functions representing the linguistic terms in the rule. Sample fuzzy rules applied in the Knowledge base is as follows:

R1: If (Behavior Problem) && (Language Disorder) && (General Mental retardation) then belongs to class Autistic

R2: If (Behavior problem Attributes (1 || 2 || …..n)) &&(Language Disorder Attributes (1 || 2 || …. n)) && (Mental retardation Attributes(1 || 2 || …..n)) then belongs to class Autistic .

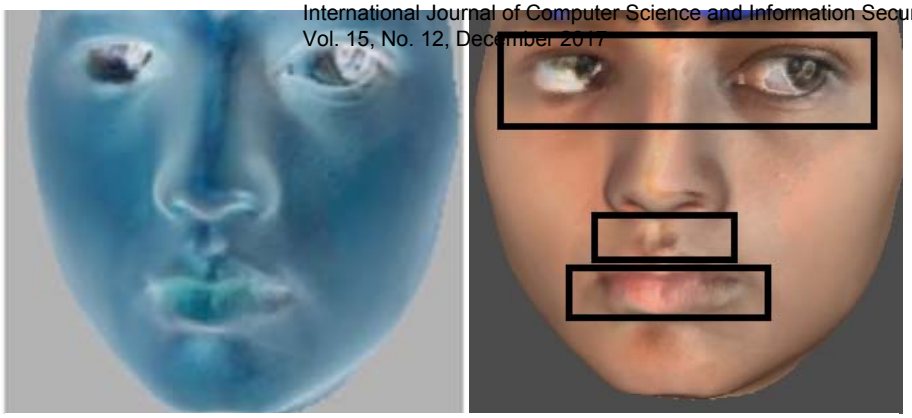
Different soft computing model have been tested like Naïve Bayes, SVM, K-Means, FCM and Neuro Fuzzy with the same input attributes using Weka tool .The performance is evaluated and the most outstanding results are shown in table4. The operational procedure of the neuro fuzzy system for autism classification is shown in figure 5

The expert system is tested and evaluated by the different stakeholders, the accuracy and evaluation survey summary is shown in figure 6 and 7.

Technique	Sample size	Inputs	Outputs	Accuracy rate
Naïve Bayes	47	12	2	100
Neuro- Fuzzy	47	22	4	98.8

Table4: Performance of Classifier.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



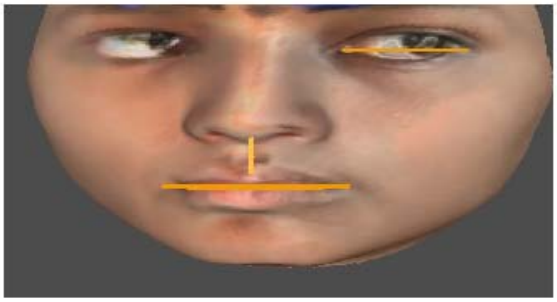


Figure 4: Facial Features

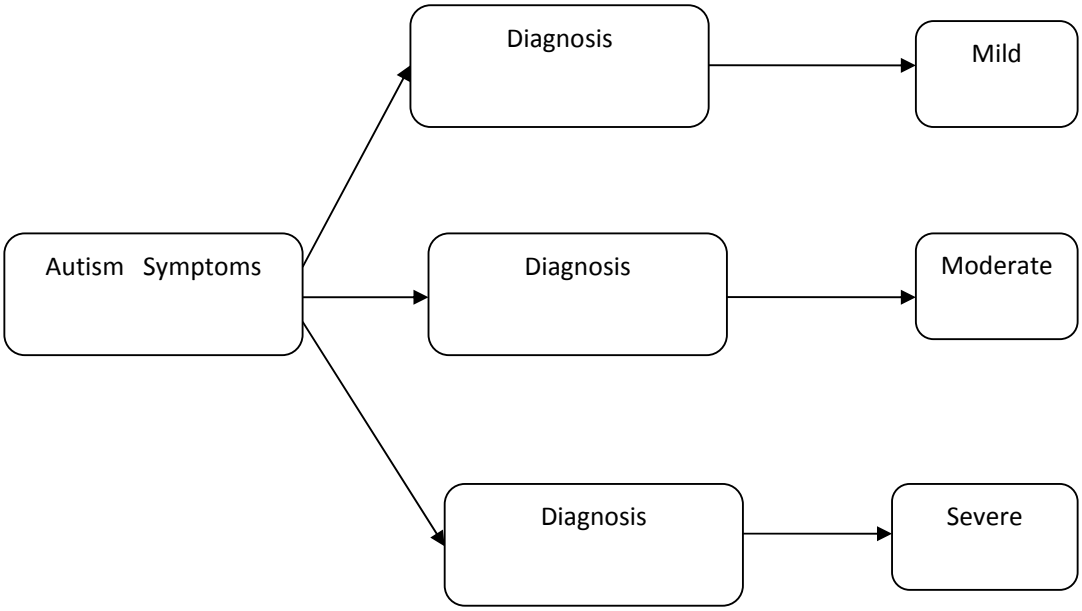


Figure 5: Operational Procedure of the Neuro-Fuzzy system for Autism classification

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

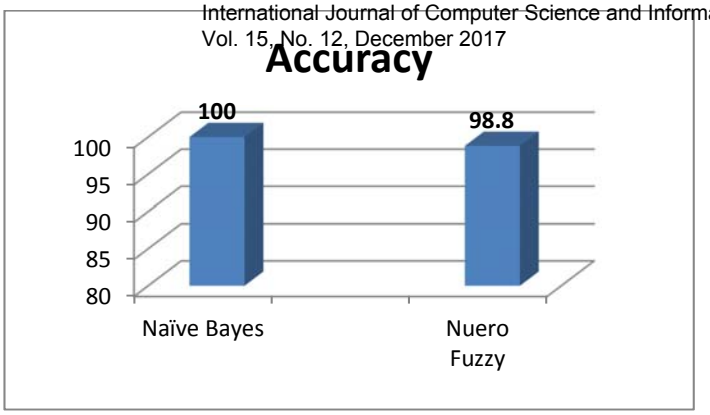


Figure: 6 Accuracy

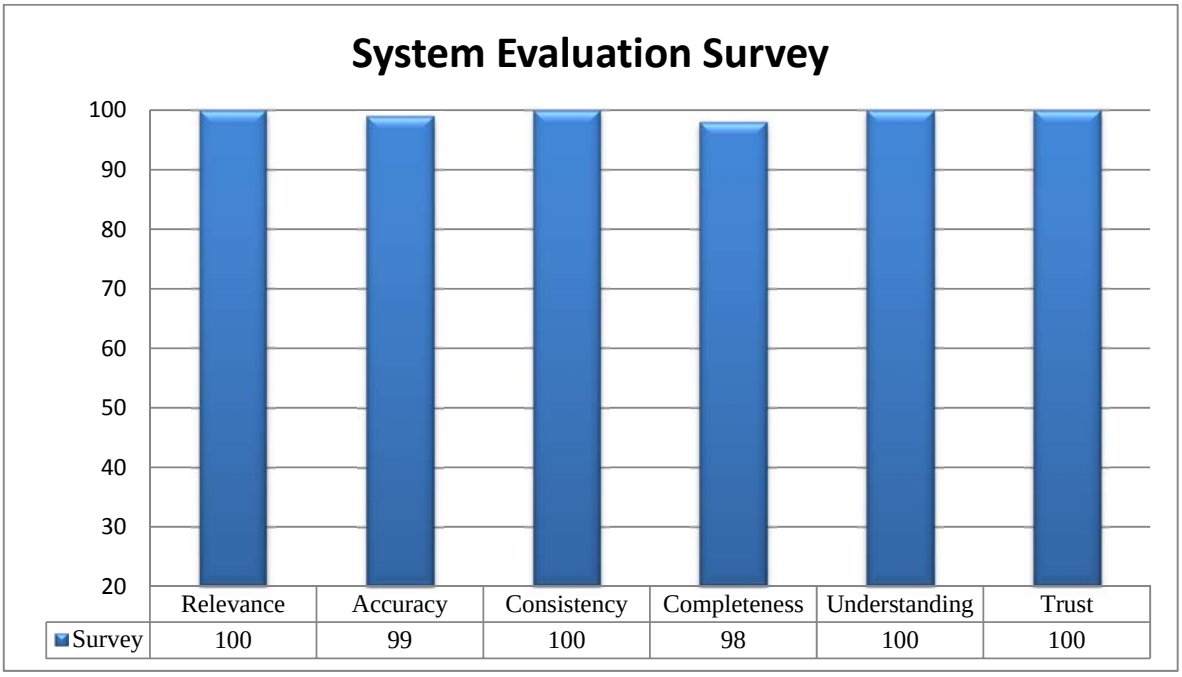


Figure 7: Expert system evaluation Survey

Conclusion: Studies related to the cause and symptoms of Autism spectrum disorder are going on around the world. Information Technology is finding lots of application in all fields. Due to the complexity and heterogeneous nature of this disorder, fewer works are reported which make use of IT in this area. Our expert system captures different inputs and produces an appropriate output. This system can be used by clinicians as a support system. The expert system is used and evaluated by 20 potential users and they all provide positive responses relating to input, output and quality of the system. Integrating 3D facial features as input to the system add a new dimension in Autism research.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

References:

[1] Myers SM, Johnson CP. Management of children with autism spectrum disorders. Pediatrics.2007;120:1162-82.

[2] Tayo Obafemi-Ajayi, Judith H. Miles, T. Nicole Takahashi, Wenchuan Qi, Kristina Aldridge, Minqi Zhang, Shi-Qing Xin, Ying He, Ye Duan. **Facial Structure Analysis Separates Autism Spectrum Disorders into Meaningful Clinical Subgroups**. *Journal of Autism and Developmental Disorders*, January 2015 DOI: 10.1007/s10803-014-2290-8

[3]. Silberberg D, Arora N, Bhutani V, Durkin M, Gulati S, Nair M, *et al.* Neuro-Developmental Disorders in India-From Epidemiology to Public Policy. Neurology. 2014; 82:P7-P324.

[4]. Mukerji S. A large scale, two phase study to estimate prevalence, and raise awareness, about autism spectrum disorders in India. Action for Autism, Jan 2009. New Delhi, India.

[5]. Myers SM, Johnson CP. Management of children with autism spectrum disorders. Pediatrics.2007;120:1162-82.

[6]. Robins DL, Casagrande K, Barton M, Chen CMA, Dumont-Mathieu T, Fein D. Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHATR/F). Pediatrics.. 2014;133;37-45.

[7]. Yasmin H. Neggers, Department of Human Nutrition, University of Alabama, Tuscaloosa AL, USA, Autism spectrum disorders: Increasing prevalence and changes in diagnostic criteria.

[8]. Vijay Sagar KJ. Research on autism spectrum disorders in India. AP J Psychol Med 2011;12 (1): 69–72.

[9]. Peter Hammond, Tim J. Hutton,1 Judith E. Allanson,Linda E. Campbell, Raoul C.M. Hennekam,Sean Holden, Michael A. Patton, Adam Shaw,I. Karen Temple, Matthew Trotter,Kieran C. Murphy, and Robin M. Winter, American Journal of Medical Genetics 126A:339–348 (2004)

[10]. Vezzetti E.; Marcolin F. (2012). 3D human face description: landmarks measures and geometrical features. In: IMAGE AND VISION COMPUTING. - ISSN 0262-8856

[11]. Gupta, S., Markey, M. K. , Bovic, A. C. (2010) “Anthropometric 3D Face Recognition”, *International Journal of Computer Vision*, Vol. 90, No. 3: 331-349.

[12]. Aldridge, K., George, I. D., Cole, K. K., Austin, J. R., Takahashi, T. N., Duan, Y., & Miles, J. H. (2011). Facial phenotypes in subgroups of prepubertal boys with autism spectrum disorders are correlated with clinical phenotypes. *Molecular Autism*, *2*, 15. http://doi.org/10.1186/2040-2392-2-15

[13]. Weigelt S, Koldewyn K, Kanwisher N (2013) Face Recognition Deficits in Autism Spectrum Disorders Are Both Domain Specific and Process Specific. PLoS ONE 8(9): e74541. doi:10.1371/journal.pone.0074541

[14]. Barbara Ruggeri ,Ugis Sarkans , Gunter Schumann ,Antonio M. Persico, Biomarkers in autism spectrum disorder: the old and the new, Springer-Verlag Berlin Heidelberg 2013

Classification of Matrix Multiplication Methods Used to Encrypt-Decrypt Color Image

Saleh A. Khawatreh

Computer Engineering Dept. Faculty of Engineering

Al-Ahliyya Amman University

skhawatreh@ammanu.edu.jo

Abstract

The digital color images are the most important types of data is now circulating on the Internet, so the protection and security of the image transition has the top priorities of computer experts. Many researchers had developed different techniques to increase the security of image transmission and most of these techniques suffer from the slow of the encryption-decryption process. In this paper we will produce a classification of the most popular encryption-decryption techniques and suggest an efficient one, the suggestion will be based in many factors such as speedup, throughput, encryption-decryption error and the hacking factor.

Key words: Encryption, decryption, speedup, throughput, hacking.

1. Introduction

Encryption is defined as the conversion of plain message (matrix which represents digital color image) into a form called a cipher text that cannot be read without decrypting the encrypted text [15]. Decryption is the reverse process of encryption which is the process of converting the encrypted text into its original plain text, so that it can be read [15]. Color image encryption is to be done before transition the image and it has to be done securely over the network so that no unauthorized user can able to decrypt the image. Image encryption, video encryption, chaos based encryption have applications in many fields including the Internet communication, transmission, medical imaging. Tele-medicine and military Communication, etc. The evolution of encryption is moving towards a future of endless possibilities. The image data have special properties such as bulk capability, high redundancy and high correlation among the pixels. Encryption techniques are very useful tools to protect secret information [3].

Encryption of data [16] has become an important way to protect data resources especially on the Internet, intranets and extranets. Encryption is the process of applying special mathematical algorithms and keys to transform digital data into cipher code before they are transmitted and decryption involves the application of mathematical algorithms and keys to get back the original data from cipher code. The main goal of security management is to provide authentication of users, integrity, accuracy and safety of data resources [16].

2. Related works

Guodong Ye [9] have presented an efficient image encryption algorithm

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

using double logistic maps, in which the digital matrix of the image is confused from row and column respectively. Confusion effect is carried out by the substitution stage and Chens system is employed to diffuse the gray value distribution. Haojiang Gao *et al*. [5] have presented an algorithm presented a Nonlinear Chaotic Algorithm (NCA) by using power and tangent functions instead of linear function. The encryption algorithm is a one-time-one-password system and is more secure than the DES algorithm. Jawahar Thakur *et al*. [17] presented a comparison between symmetric key algorithms such as DES, AES, and Blowfish. The parameters such as speed, block size, and key size are considered to evaluate the performance when different data loads are used. Blowfish has a better performance than other encryption algorithms and AES showed poor performance results compared to other algorithms due to more processing power.

Khaled Loukhaoukha *et al.* [9] introduced an image encryption algorithm based on Rubik's cube principle. The original image is scrambled using the principle of Rubik's cube and then XOR operator is applied to rows and columns of the scrambled image using two secret keys. Liu Hongjun *et al*. [18] designed a stream-cipher algorithm based on one-time keys and robust chaotic maps. The method uses a piecewise linear chaotic map as the generator of a pseudo-random key stream sequence.

M. Zeghid *et al*. [19] analyzed the AES algorithm, and added a key stream generator (A5/1, W7) to AES to ensure improved encryption performance mainly for the images. The method overcomes the problem of textured zones existing in other known encryption algorithms. Maniccam *el al.* [20] presented a method for image and video encryption and the encryption methods are based on the SCAN methodology. The image encryption is performed by SCAN-based permutation of pixels and a substitution rule which together form an iterated product cipher. The pixel rearrangement is done by scanning keys and the pixel values are changed by substitution mechanism. Figure 1 shows the basic SCAN patterns used in [16]. Mohammad Ali el al. [21] introduced a block-based transformation algorithm based on the combination of image transformation and the Blowfish algorithm. The algorithm resulted in the best performance by the lowest correlation and the highest entropy. The characteristics of AES are its security and resistance against attacks and the major characteristic of RC4 algorithm is its speed [11]. A hybrid cipher by combining the characteristics of AES and RC4 is developed and 20% improvement in speed is achieved when compared to the original AES and a higher security compared to the original RC4 [13].

Rizvi *et al.* [12] analyzed the security issues of two symmetric cryptographic algorithms Blowfish and CAST algorithm and then compared the efficiency for encrypting text, image, and audio with the AES algorithm across different widely used Operating Systems. For text data, all algorithms run faster on Windows XP but Blowfish is the most efficient and CAST run slower than AES. Blowfish encrypts images most efficiently on all the three

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

platforms. For audio files, CAST performs better than Blowfish and AES on Windows XP but on Windows Vista and Windows 7, there is no significant difference in the performance of CAST and AES; however, Blowfish encrypts audio files at less speed.

Sanfu Wang *et al*. [21] presented an image scrambling method based on folding transform to folding matrix which is orthogonal and enables to fold images either up-down or left-right. When an image is folded this way repeatedly, it becomes scrambled. The scrambling algorithm has an effective hiding ability with small computation burdens as well as wide adaptability to images with different scales.

Sathishkumar G.A *et al*. [14] presented a pixel shuffling, base 64 encoding based algorithm which is a combination of block permutation, pixel permutation, and value transformation. The crypto system uses a simple chaotic map for key generation and a logistic map was used to generate a pseudo random bit sequence. The total key length is 512 bits for each round and the key space is approximately 2512 for ten rounds. Shao Liping *et al.* [4] proposed a scrambling algorithm based on random shuffling strategy which could scramble non equilateral images and has a low cost to build coordinate shifting path. The algorithm is based on permuting pixel coordinates and it could be used to scramble or recover image in real time. T. Sivakumar, and R. Venkatesan [4] proposed a novel image encryption approach using matrix reordering this approach was tested and some comparisons with other techniques were done.

Ziad A. Alqadi and others in [1] and [2] have presented a technique using direct and inverse conversions to convert a color image to gray image and vice versa, this technique can be useful to be used in color image encryption decryption.

3. Propped methods

3-1 First method (proposed 1): Using each of the components of the color image

This method for encryption can be implemented in the following steps:

1. Get the original color image.
2. Extract the red, green, and blue matrices from the original color image (each of them is 2 dimensional matrix),
3. Reshape each matrix in step 2 to square matrix.
4. Generate one random square matrix for each component to be used as a private key.
5. Encrypt each component by applying matrix multiplication of the matrix component and its private key.
6. Reshape each encrypted matrix to its original size.
7. Form the encrypted color image.

The decryption phase can by implemented applying the following steps:

1. Get the decrypted color image.
2. Extract the red, green, and blue matrices from the original color

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

image (each of them is 2 dimensional matrix),

3. Reshape each matrix in step 2 to square matrix.
4. Use each private key
5. Decrypt each component by applying matrix multiplication of the matrix component and the inverse it's private key.
6. Reshape each decrypted matrix to its original size.
7. Form the decrypted color image.

The following matlab code was written to implement this method

```
clear all
close all
a=imread('C:\Users\User\Desktop\flower-color-combinations.jpg');
subplot(2,2,1)
imshow(a), title 'Original image'
subplot(2,2,2)
imhist(a(:,:,1)), title 'Red component histogram'
subplot(2,2,3)
imhist(a(:,:,2)), title 'Green component histogram'
subplot(2,2,4)
imhist(a(:,:,3)), title 'Blue component histogram'
tic
b1=a(:,:,1);
b2=a(:,:,2);
b3=a(:,:,3);
b1=reshape(b1,200*300,1);
b2=reshape(b2,200*300,1);
b3=reshape(b3,200*300,1);
for i=60001:60025
   b1(i,1)=0;
   b2(i,1)=0;
   b3(i,1)=0;
end
c1=reshape(b1,245,245);
c2=reshape(b2,245,245);
c3=reshape(b3,245,245);
k1=rand(245,245);
k2=rand(245,245);
k3=rand(245,245);
c1=double(c1);
c2=double(c2);
c3=double(c3);
e1=c1*k1;
e2=c2*k2;
e3=c3*k3;
toc
tic
d1=e1*inv(k1);
d2=e2*inv(k2);
d3=e3*inv(k3);
d11=reshape(d1,245*245,1);
d12=reshape(d2,245*245,1);
d13=reshape(d3,245*245,1);
for i=1:60000
   d21(i,1)=d11(i,1);
   d22(i,1)=d12(i,1);
   d23(i,1)=d13(i,1);
end
d31=uint8(d21);
d32=uint8(d22);
d33=uint8(d23);
d41=reshape(d31,200,300);
d42=reshape(d32,200,300);
d43=reshape(d33,200,300);
d4(:,:,1)=d41;
d4(:,:,2)=d42;
d4(:,:,3)=d43;
toc
figure
subplot(2,2,1)
imshow(d4), title 'Decrypted image'
subplot(2,2,2)
imhist(d4(:,:,1)), title 'Decrypted red component histogram'
subplot(2,2,3)
imhist(d4(:,:,2)), title 'Decrypted green component histogram'
subplot(2,2,4)
imhist(d4(:,:,3)), title 'Decrypted blue component histogram'
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3-2 Second method (proposed 1): Converting color image to 2 dimensional matrix

The encryption phase here is consisted of the following steps:

1. Get the original digital color image as a 3 dimensional matrix(m).
2. Reshape m into 1 column matrix(r).
3. Get the size of r (s).
4. If s is a square number proceed to step 6.
5. Find the nearest square number to s and adjust s to this number, adjust r by padding zeros.
6. Reshape r to square matrix (r1).
7. Generate a double random square matrix with size equal r1 size, this matrix will be used as a private key for encryption-decryption (k).
8. Save k to be used in the decryption phase.
9. Get the encrypted image (e) by applying matrix multiplication of r1 and k.
10. Reshape e into 1 column matrix (e1).
11. Omit the padded zeros from e1.
12. Reshape e1 into 3 dimensional matrix to get the encrypted color image.

The decryption phase can be implemented applying the following steps:

1. Get the encrypted digital color image as a 3 dimensional matrix(en1).
2. Reshape en into 1 column matrix(en2).
3. Get the size of en2 (s).
4. If s is a square number proceed to step 6.
5. Find the nearest square number to s and adjust s to this number, adjust en2 by padding zeros.
6. Reshape en2 to square matrix (en3).
7. Use the private key k.
8. Get the decrypted image (di) by applying matrix multiplication of r1 and the inverse of k.
9. Reshape di into 1 column matrix (di1).
10. Omit the padded zeros from di1.
11. Reshape di1 into 3 dimensional matrix to get the decrypted original color image.

The following matlab code was written to implement this method

```
clear all
close all
a=imread('C:\Users\User\Desktop\flower-color-combinations.jpg');
subplot(2,2,1)
imshow(a), title 'Original image'
subplot(2,2,2)
imhist(a(:,:,1)), title 'Red component histogram'
subplot(2,2,3)
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
imhist(a(:,:,2)), title 'Green component histogram'
subplot(2,2,4)
imhist(a(:,:,3)), title 'Blue component histogram'
tic
b=reshape(a,200*300*3,1);
for i=180001:180625
   b(i,1)=0;
end
c=reshape(b,425,425);
k=rand(425,425);
c=double(c);
e=c*k;
toc
tic
d=e*inv(k);
d1=reshape(d,425*425,1);
for i=1:180000
   d2(i,1)=d1(i,1);
end
d3=uint8(d2);
d4=reshape(d3,200,300,3);
toc
figure
subplot(2,2,1)
imshow(d4), title 'Decrypted image'
subplot(2,2,2)
imhist(d4(:,:,1)), title 'Decrypted red component histogram'
subplot(2,2,3)
imhist(d4(:,:,2)), title 'Decrypted green component histogram'
subplot(2,2,4)
imhist(d4(:,:,3)), title 'Decrypted blue component histogram'
```

3-3 Third method (proposed 1):Converting color image to Gray image

This method can be implemented as first method but the color image is to converted to gray image using direct conversion proposed by the author in [1], then the gray image can be encrypted as in method 1, after that the encrypted gray image can be decrypted and converted to color image using the inverse conversion mentioned in [1].

4. Experimental results

The proposed methods were implemented several times using different color images with different sizes and the results always give a correlation coefficient equal 1 between the original image and the decrypted one, which means that the methods are 100% correct and do not lead to any damage of information, figure 1 and 2 show the original image and the decrypted one with the histogram of each component of the color image.

The proposed method is also very secure and it is implausible to hack the image because the private key has the following features:

- Private key is a 2 dimensional matrix with a huge size.
- Each element in the private key is a random double number which make it impossible to guess.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

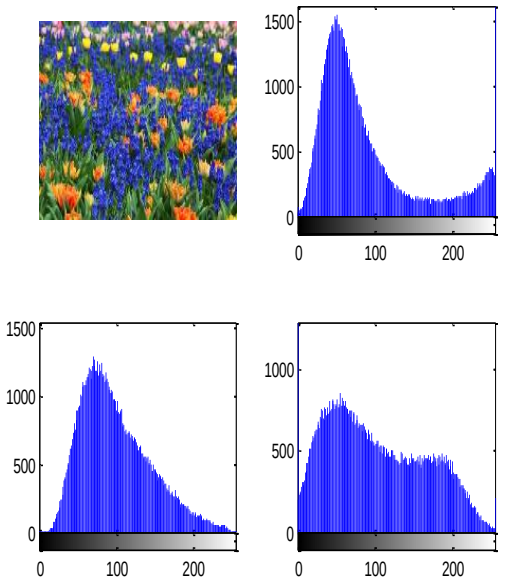


Figure 1: sample of the original color image

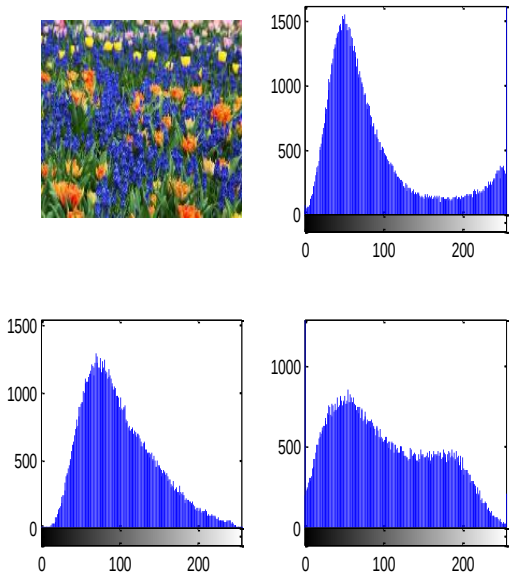


Figure 2: Decrypted color image.

The encryption and decryption times were calculated and compared with other methods mentioned in the related works, these results are listed in table 1.

Table 1: Comparisons results

Method	Direct conversion time (s)	Inverse conversion time (s)	Encryption time (s)	Decryption time(s)	Total time(s)	Speed up	Throughput(MB/s)
Proposed 1	0	0	0.006207	0.067798	0.07400	1	21.2549
Proposed 2	0	0	0.027985	0.156311	0.18433	2.4905	8.53 43
Ref[1] HSI	0.078	0.032	0.02032	0.02541	0.1557	2.1041	10.1019
REF [1]R'G'B'	0.015	0.015	0.02032	0.02541	0.0757	1.0230	20.7776
Ref.[4]	0	0	0.23	0.23	0.46	6.2162	3.41 93
Ref.[5]	0	0	0.5	0.5	1.0	13.5135	1.57 29
Ref.[6]	0	0	0.12	0.12	0.24	3.2432	6.55 36
Ref.[7], (A-I)	0	0	0.56	0.56	1.12	15.1351	1.40 43
Ref.[7],(A-II)	0	0	1.01	1.01	2.02	27.2973	0.77 86
Ref.[8]	0	0	0.4	0.4	0.8	10.8108	1.96 61

The speedup was calculated by dividing the total time of the method by the total time of proposed 1 (which was taken as a reference because it has the best efficiency).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The throughput was calculated by dividing the color image size by the total time.

For clarity we can represent the data in table 1 by figures 3 and 4.

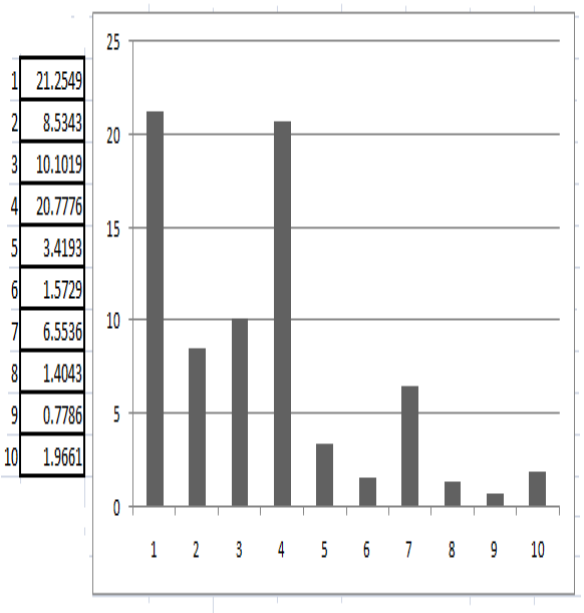


Figure 3: Methods throughput

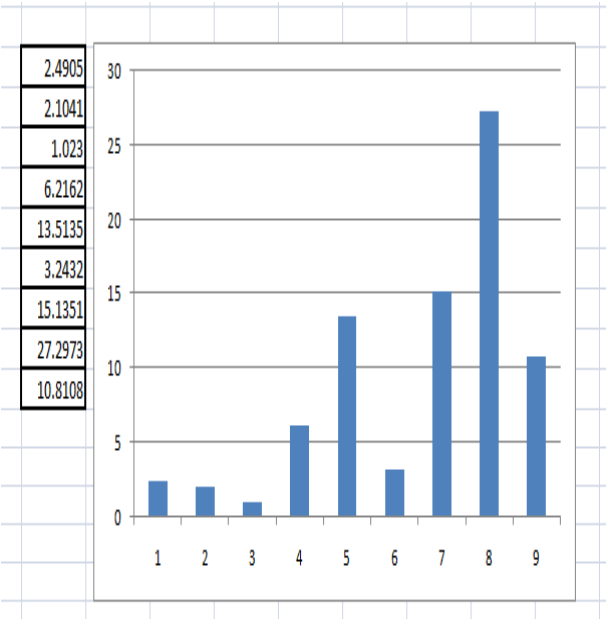


Figure 4: Speedup of the method reference to proposal 1

From the above results we can see that the proposal method 1 has the best efficiency.

Conclusions

A methods of encryption-decryption of color image were proposed and a survey analysis was done and it was shown that proposed 1 method has the best performance because it characterized with following features:

- **Best speed in encryption phase.**
- **Best speed in decryption phase.**
- **Best throughput.**
- **No any damage of information.**
- **Impossible to hack.**

References

[1]: Majed O. Al-Dwairi, Ziad A. Alqadi, Amjad A. AbuJazar and Rushdi Abu Zneit, Optimized True-Color Image Processing, World Applied Sciences Journal 8 (10): 1175-1182, 2010 ISSN 1818-4952.

[2]: Waheeb, A. and Ziad AlQadi, 2009. Gray image reconstruction. Eur. J. Sci. Res., 27: 167-173.

[3]: Rojo, M.G., G.B. García, C.P. Mateos, J.G. García and M.C. Vicente, 2006. Critical comparison of 31 commercially available digital slide systems in pathology. Int. J. Surg. Pathol., 14: 285-305.

[4]: T. Sivakumar, and R.Venkatesan , A Novel Image Encryption Approach using Matrix Reordering, WSEAS TRANSACTIONS on COMPUTERS,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Issue 11, Volume 12, November 2013,pp 407-418.

[5]: Haojiang Gao, Yisheng Zhang, Shuyun Liang and Dequn Li, "A New Chaotic Algorithm for Image Encryption", *Elsevier Science Direct*, vol. 29, no. 2, 2006, pp.393-399.

[6]: Khaled Loukhaoukha, Jean-Yves Chouinard, and Abdellah Berdai, "A Secure Image Encryption Algorithm Based on Rubik's Cube Principle", *Journal of Electrical and Computer Engineering*, 2011, pp. pp.1-13.

[7]: Xiaomin Wang, and Jiashu Zhang, "An Image Scrambling Encryption using Chaos- controlled Poker Shuffle Operation", *IEEE International Symposium on Biometrics and Security Technologies*, Islamabad, 23-24 April 2008, pp.1-6.

[8]: G. Chen, Y. Mao, and C. K. Chui, "A Symmetric Image Encryption Scheme based on 3D Chaotic Cat Maps", *Chaos, Solitons and Fractals*, Vol. 21, No. 3, 2004, pp.749–761.

[9]: Guodong Ye, "An Efficient Image Encryption Scheme based on Logistic maps", *International Journal of Pure and Applied Mathematics*, Vol.55, No.1,2009, pp. 37-47.

[10] Han Shuihua and Yang Shuangyuan, "An Asymmetric Image Encryption Based on Matrix Transformation", *ECTI Transactions on Computer and Information Technology*, Vol.1, No.2, 2005, pp. pp.126-133.

[11] Prabhudesai Keval Ketan and Vijayarajan V, "An Amalgam Approach using AES and RC4 Algorithms for Encryption and Decryption", *International Journal of Computer Applications*, Vol.54, No.12, 2012, pp.29-36.

[12]: S.A.M Rizvi, Syed Zeeshan Hussain and Neeta Wadhwa, "A Comparative Study of Two Symmetric Encryption Algorithms Across Different Platforms", *International Conference on Security and Management (SAM'11)*, World Academy of Science, USA, 2011.

[13] Sanfu Wang, Yuying Zheng and Zhongshe Gao, "A New Image Scrambling Method through Folding Transform", *IEEE International Conference on Computer Application and System Modeling*, Taiyuan, 22-24 Oct. 2010, pp. v2-395-399.

[14] G.A. Sathishkumar and K. Bhoopathy Bagan, "A Novel Image Encryption Algorithm Using Pixel Shuffling and BASE 64 Encoding Based Chaotic Block Cipher, *WSEAS Transactions on Computers*, Vol.10, No. 6, 2011, pp. 169-178.

[15] John Justin M, Manimurugan S, "A Survey on Various Encryption Techniques", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[16] Ephin M, Judy Ann Joy and N. A. Vasanthi, "Survey of Chaos based Image Encryption and Decryption Techniques", Amrita International Conference of Women in Computing (AICWIC'13) Proceedings published by International Journal of Computer Applications (IJCA).

[17]: Jawahar Thakur, and Nagesh Kumar, "DES, AES, and Blowfish:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Symmetric Key Cryptography Algorithms Simulation Based Performance Analysis", *International Journal of Emerging Technology and Advanced Engineering*, Vol.1, No.2, 2011, pp.6-12.

[18]: Liu Hongjun and Wang Xingyuan, "Color image encryption based on one-time keys and robust chaotic maps", *Journal of Computers and Mathematics with Applications (Elsevier)*, Vol.59, 2010, pp. 3320-3327.

[19]: M. Zeghid, M. Machhout, L. Khriji, A. Baganne, and R. Tourki, "A Modified AES Based Algorithm for Image Encryption", *World Academy of Science, Engineering and Technology*, Vol.3, 2007, pp.526-531.

[20]: S.S. Maniccam, and N.G. Bourbakis "Image and Video Encryption using SCAN Patterns", *The Journal of the Pattern Recognition Society*, Vol.37, 2004, pp.725-737.

[21]: Mohammad Ali Bani Younes and Aman Jantan, "Image Encryption using Block-Based Transformation Algorithm", *IAENG International Journal of Computer Science*, Vol.35, No.1, 2008, pp.3-11.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Document Processing and Arabic Optical Character Recognition: A User Perspective

Yasser M. Alginahi, Senior Member, *IEEE*

***Abstract*— The technological era we live in has influenced our daily activities at home, work and everywhere. The software technology has prospered, and new technology is being introduced to end-users every day. Document image analysis software has been influenced with more applications being developed. Hence, this inspired us to perform a general study on application software with emphasis on document analysis software i.e., Optical Character Recognition (OCR). This study starts with a brief review on recent works on document processing and Arabic OCR. Following this, a questionnaire survey is conducted to investigate the capability and familiarity of individuals with the top four Arabic OCR software, in order to provide recommendations and future research directions in this area. The results show that OCR is an essential technology which should be available with operating system tools. From the survey results, many of the respondents are not familiar with the top four commercial OCR, even those who are in the Information Technology (IT) sector. The study concludes that the available commercial OCR software are becoming more efficient; however, the accuracy rate reported needs to be further evaluated to provide a more accurate performance/recognition rate and, further investigation is needed to analyze the reported commercial accuracy rates. Finally, the study concludes with recommendations and future research directions.**

***Index Terms*— Character recognition, Document Analysis, Editing Software, Optical Character Recognition Software, Scanner.**

I. Introduction

DOCUMENT processing is an area of research that includes pattern recognition, artificial intelligence, data mining, information retrieval, image processing, and computer vision. Document processing, i.e. OCR, has been used in coordinating and conducting business transactions, mail sorting, check processing, passport processing, online publishing, digital libraries … etc. The advancements in technology has revolutionized document processing, with the increase in speed, emergence of new storage medium, and increase storage capacity. Document processing is the capture of information from a paper medium into a digital medium, in other words, digitization. This may involve processing documents containing text only or documents containing mixed content (text and images). The process starts by separating/segmenting the documents into text and images which are then processed separately. Nowadays, with the available advanced technology, documents can be processed automatically with high precision results. Text processing involves the conversion of paper or image documents into digital formats, i.e., OCR, which can be further processed. OCR is a common method of digitizing printed texts, to enable searching, editing, storing, transmitting or further processing of documents [1]. Similarly, image processing involves editing, storing, transmitting and further processing of images.

The term digitization has emerged since the last two decades. It is the conversion of paper documents into electronic formats, and nowadays the terms e-government, e-library, e-services, e-learning, e-commerce … etc. have emerged; meaning digitization is being applied in many sectors/areas. Hence, the reduction in the amount of papers used is seen in many areas. Many libraries are reducing the paper books on their shelves and converting to digital libraries. Therefore, with the vast amount of information to be digitized many governments, firms and libraries around the world have started projects to digitize their paper contents, especially ancient manuscripts, thesis, books, and old documents. Therefore, digital libraries are now available online, which makes it easier to find information on a click of a button [2].

Computers are becoming faster and more reliable in document processing; in addition, database technologies for archived information maintenance have existed for long time and been used to store vast amount of information. Thus, OCR intelligence can be applied to digital image documents, which means the reduction in cost and time, since searching for information on a computer is much faster than finding it in a pile of dusty documents, which may take from several minutes to several hours. Therefore, the purpose of digitization is to produce digital documents which can be edited, stored, searched, transmitted online and used in other applications such as machine translation, text mining, and text-to-speech conversion.

This study explores the user experience with application software, document processing software and the four top available Arabic OCR software. The purpose is to get the user's perspective on document processing software. The main objectives of this study are:

- To explore the level of competence of individuals with basic application software skills in creating and sharing documents.
- To investigate how familiar and capable individuals are with OCR software.
- To investigate the accuracy of the recognition rate of the top four commercial OCR software from the user perspective.
- To study the current status of research in the area of Arabic OCR.

This paper is organized as follows. After this introduction, Section 2 provides the literature survey. Section 3 explains the research methodology. Section 4 presents the results and analysis. Section 5 provides the limitations and discussion, and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

finally the conclusion and future work are stated in section 6.

II. Literature Review

The efficiency of document analysis has progressed and advanced with technology. Applications which use OCR have increased with the emergence of more efficient OCR software. The early versions of OCR software were very limited and worked for specific applications on specific fonts at a time. Early OCR may be traced to technologies involving telegraphy and creating reading devices for the blind [3], such as the machine developed by Emmanuel Goldberg in 1912 which converts characters into standard telegraph code and the Optophone device developed by Fournier D'Albe in 1914. Developments in OCR continued from the early 20th century until the era of the first digital computer in the early 1940s. Then, in the 1950s OCR became part of the business world [4]. Nowadays, advanced systems can produce a high degree of recognition accuracy for most fonts and different file formats, providing results which are similar to the format and layout of the input documents including columns, images, fonts, styles … etc. Currently, OCR is available online as a service in a cloud-computing environment i.e., APIs, which can be used from any device connected to the Internet. These OCR API provide a simple way of parsing images and even multi-page documents to provide the results as a text file. In addition, various commercial and open source OCR systems are available for most of the common languages [5 – 6].

The work on document analysis is very active. Recently, the shift in research has moved from segmenting simple documents with text and images to segmenting text from scenes, billiard boards, movies … etc. Document analysis goes through many steps and starts with processing the document digital image and ends with characters being recognized. The document analysis starts with a digital document image, then preprocessing is applied depending on the quality of the image. Preprocessing may include filtering noise, de-skew image, converting to gray or binary image … etc. Next, page segmentation techniques are applied to identify different regions on the page, this is shown in Figure 1. After this, the text and graphics/images regions are both processed separately. Processing of the text portion of a document image includes: script recognition, font recognition, line/paragraph segmentation and word segmentation, Figure 2. Further processing of word segmentation includes, Holistic Word Recognition, Integrated Segmentation and Recognition, and Character Segmentation (OCR), Figure 3.

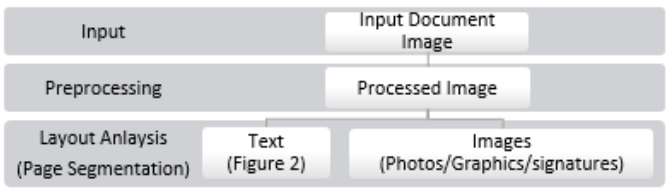


Figure 1: Document Image Analysis (Page Segmentation)

A general OCR system, Figure 3, goes through the following stages: preprocessing, character segmentation, feature extraction, classification and recognition. After a digital image is fed into the system, preprocessing techniques may be used to remove noise, de-skew image, sharpen images … etc.

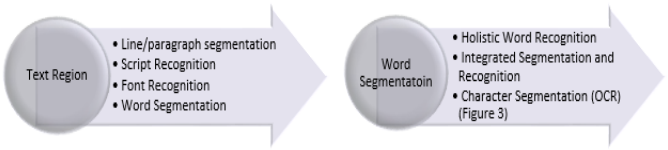


Figure 2: Text Region Processing

Next, the characters are isolated or segmented, and normalized. In addition, other techniques may be used to help prepare the characters for the feature extraction stage. The next stage is to classify and recognize the characters. Finally, post processing (contextual, grammatical information or data dictionary authentication), may be used to aid in achieving higher recognition rates. All these stages work in a pipeline fashion. The previous stage feeds into the next stage; therefore, the success of each stage guarantees an efficient OCR system.

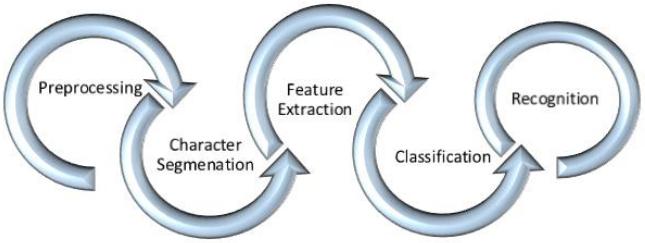


Figure 3: General OCR System

The work on OCR is still active in all stages presented in Figure 3. New techniques for OCR preprocessing, feature extraction, classification, and recognition are being published in recent literature. The work in [7] presents a comprehensive survey on character segmentation and challenges for segmentation of Arabic script based languages, i.e, Arabic, Urdu, Persian, Pashto, Sindi and Malay (Jawi). However, this research concentrated more on Urdu and concluded that the cursive nature of the Arabic script is the main challenge in character segmentation especially in Nasta'liq writing style compared to Naskh. Some other works related to document analysis include analysis of documents with non-uniform background [8], page segmentation [9], Arabic script recognition [10] [11], and segmentation of Arabic characters [12]. In addition, recent surveys in the areas of printed and handwritten Arabic and Urdu OCR research are available in the literature [13 – 17].

The research work on OCR for cursive scripts, such as Arabic, are not as mature as is the case with Latin scripts. Therefore, document image databases are needed for training, testing, validating and eliminating of errors. The work in [18] presents a multilingual image database created from various texts containing multiple fonts collected from various sources and 84 language scripts including Arabic script. The Arabic script is adopted by languages such as Kashmiri, Kurdish, Pashto, Persian, Punjabi (Shahmukhi), Sindhi, Urdu, and Uyghur. The images in the database were converted into word,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

single line, multiline and paragraph images. Therefore, this database provides a platform for a step towards the establishment of standardized image database for the OCRs of the world language scripts. This database can be expanded to include more images and scripts, which researchers in the area of document analysis can benefit from [18]. Other works which discussed databases include the following scripts: Sindi [19], handwritten Persian [20], Pashto [21], handwritten Urdu [22], and Arabic [23]

Recent OCR works on printed Arabic scripts include the works done on the following Arabic scripts Urdu [24], Pashto [25], Persian [26], Punjabi [27], Uyghur [28], and Sindhi [29]; and for handwritten Arabic scripts include Urdu [30], Pashto [31], Persian [32], Punjabi [33], Uyghur [34], and Sindhi [35].

The research in the area of printed Arabic OCR resulted in a number of high quality commercial recognition systems. However, the general problems of such systems have not been solved yet. The main challenges, which are not solved in many OCR systems, still linger: multi-language/script recognition, multi-column reading order, very noisy documents, layout retention, and handwriting script [36]. The authors in [36] developed a multi-stage approach to document analysis involving preprocessing, content segmentation, recognition, and correction.

Currently, commercial OCR software for printed Latin scripts have reported 100% accuracy rates; however, this rate is not possible without the use of contextual, grammatical information or data dictionary authentication even where clear imaging is available. For example, using a smaller dictionary can help achieve a higher recognition rate for reading the amount line of a cheque. As a result, the recognizing of words from a dictionary is easier than trying to parse individual characters from script [37]. One study based on recognition of newspaper pages from the 19th and early 20th century concluded that character-by-character OCR accuracy for commercial OCR software varied from 81% to 99% [38]. On the other hand, the accuracy rates for hand-printed text is 80% to 90% on neat, clean hand-printed characters, and hence this accuracy rate still translates to dozens of errors per page, making the technology useful only in very limited applications. Furthermore, the accuracy rates for cursive text is still low especially in scripts other than Latin where more research is needed to achieve higher rates. Recognition of hand-printing, cursive handwriting, and printed text in other scripts are still the subject of active research [37 – 39].

OCR is used to be part of scanner software; then, it was available to be used on image documents without the use of a scanner. Now, OCR software is available as mobile apps where documents are being processed by taking an image using the mobile camera. Also, some OCR companies are providing OCR as an online service, where the documents are being processed on the cloud, i.e., OCR API [6]. Therefore, OCR applications are real-time and are used in government offices for scanning passport applications, license plate identification, … etc. Nowadays, several OCR companies provide online service to process documents on cloud [6]. In this study, the top four Arabic commercial OCR software are studied in order to provide recommendations and future directions in the area of OCR; these software are, Adobe Acrobat, OmniPage Standard, ABBYY FineReader, and Readiris. Table 1 shows the OCR performance rates and number of languages supported for these software. The data in Table 1 is obtained from www.toptenreviews.com [40].

Table 1: Reported OCR performance rates and number of languages supported

	Usability	Text Accuracy	Languages Recognized
Adobe Acrobat	91%	100%	190
OmniPage Standard	83%	99.80%	120
ABBYY FineReader	66%	99.84%	190
Readiris	75%	99.83%	130

*source: http://www.toptenreviews.com/business/software/best-ocr-software/ [40]

Most of these software report perfect or close to perfect accuracy rates; however, no specific accuracy rates are provided for each of the languages supported by the software. Therefore, further studies are needed to evaluate the recognition rates of each language. As it can be seen, the highest number of languages supported is 190 by three software namely: Adobe Acrobat, and ABBYY FineReader.

A simple search for the term "OCR" on www.download.com provides 503 hits as of 24th of April 2017, between open source and commercial software [41]. Most of these software are open sources and they are usually research projects implemented by individuals or research groups. In general, the performance of open source OCR software is usually very low, have few editing capabilities, and is far from reaching the performance levels of commercial OCR software [41].

III. Research Methodology

A semi-structured questionnaire survey was designed. The survey aims to investigate the objectives outlined in the previous section. The questionnaire was first prepared and then disseminated to five subjects. After that, the questionnaire was modified. The survey questions are given in Table 2. Then three semi-structured interviews with faculty members from the IT sector were conducted to discuss the survey questions. The purpose of this step is to get feedback and comments on the clarity of the survey in order to consolidate some of the findings that were observed from the survey. Then, after an in-depth review of the subject matter, the semi-structured questionnaire was designed using www.monkeysurvey as a tool for data collection.

The survey was disseminated to the author's contacts using email, WhatsApp, and Researchgate contacts. These individuals are mainly working in higher-level academic institutions all over the world. Approximately, 800 emails were sent, and it is expected that the response rate will be between 10 – 20%. This study explores the user experience with application software, document processing software and the four top Arabic OCR software in order to investigate the user's perspective on document analysis software.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 2: The survey questions with no. of responses to each question

No.	Question Statement	Type of Question	Answered	Skipped	Percentage Answered
1	In what country do you live?*	Dropdown	124	2	98.4%
2	What is your gender?	Multiple Choice	124	2	98.4%
3	What is your age?	Dropdown	126	0	100.0%
4	What is your area of expertise/study?*	Dropdown	125	1	99.2%
5	What is the highest level of education you have completed?	Dropdown	125	1	99.2%
6	How important to your learning is it to have access to technology?	Multiple Choice	125	1	99.2%
7	How capable are you with each of the following application software to create and share documents?	Matrix/ rating scale	125	1	99.2%
8	How often do you use the following application software?	Matrix/ rating scale	125	1	99.2%
9	Which of the following problems you have encountered with the application software in Question 8? *	Multiple Choice	106	20	84.1%
10	Have you used a scanner or have you taken a photo/image of a text document (which may also contain images/photos)?	Yes/No	118	8	94.9%
11	Have you used an OCR software to convert paper documents into editable text files?	Yes/No	120	7	95.2%
12	For which of the following languages have you used OCR software? *	Multiple Choice	102	24	81.0%
13	How often have you used the following OCR software?	Matrix/ rating scale	114	12	90.5%
14	How capable are you with each of the following OCR Software?	Matrix/ rating scale	108	18	85.7%
15	Which of the following problems have you encountered with the OCR Software in Question 14? *	Multiple Choice	90	36	71.4%
16	Do you know that the following features are available with many OCR Software?	Matrix/ rating scale	110	16	87.3%
17	How do you rank the accuracy of OCR in the following language (If applicable)?*	Matrix/ rating scale	110	16	87.3%
18	Have you used OCR software to convert scanned handwritten text documents into an editable text document?	Yes/No	111	15	88.1%
19	If your answer to Question 18 is yes, How was the accuracy? And in which language? *	Matrix/ rating scale	69	57	54.8%
20	Please select the statement that indicates how you feel about OCR? *	Multiple Choice	106	20	84.1%
21	Comments/Suggestions/concern or anything you want to add to the survey.	Comment box	22	104	17.5%

* questions which include an "other" option

The survey questionnaire starts by collecting information on the respondent's profile, such as gender, age, country of residence, education level, and area of expertise. The second part of the survey was divided into two parts: application software technologies and OCR software. The type of questions were mainly multiple choice, yes/no, dropdown, matrix/rating scale, and comment box. In addition, some questions had an "other" option giving the participants the flexibility to provide their own answers instead of choosing from the provided ones. The survey was conducted between January 18, 2017 and February 28, 2017. The questions used in the survey are provided in Table 2. The total number of responses were 126, approximately 15.8% of the number of emails sent, and this rate is an acceptable response rate and according to the projected expectations. During the survey period, the author sent several reminders, by email and WhatsApp to the contacts; however, this did not help much to remind people to answer the survey, and only very few responses were seen added to the number of respondents after each reminder. The low response rate could be because most of the contacts are Doctorate and Master level individuals, and they are usually too busy to spare 10 – 15 minutes to fill the survey.

IV. RESULTS AND ANALYSIS

The survey is divided into three parts: participants personal profile information, such as gender, age, country of residence, education level and area of expertise; application software technologies; and OCR software. In this section, the questionnaire results are presented and analyzed.

A. Personal Profile Information

The total number of responses were 126, (69.4% males and 30.6% females), with two respondents who did not indicate their gender. The respondents were from 25 countries and covered the five continents, Table 3. The ages of respondents are given in Table 4. It is observed that approximately 64% of the respondents are between the ages 30 and 50, and 20% below 30 years of age. This shows that usually individuals who are employed and educated may use such software.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 3: The number of respondents across the continents

No.	Country	Percent Reponses	Response count	Continent	Percent Response
1	Algeria	4.0%	5	Africa	20.2%
2	Egypt	9.7%	12		
3	Ethiopia	0.8%	1		
4	Ghana	0.8%	1		
5	Morocco	0.8%	1		
6	Nigeria	0.8%	1		
7	Sudan	2.4%	3		
8	Tunisia	0.8%	1		
9	Bahrain	0.8%	1	Asia	65.3%
10	India	4.8%	6		
11	Japan	0.8%	1		
12	Jordan	2.4%	3		
13	Malaysia	8.1%	10		
14	Oman	0.8%	1		
15	Pakistan	3.2%	4		
16	Palestine	0.8%	1		
17	Qatar	0.8%	1		
18	Saudi Arabia	41.1%	51		
19	Yemen	1.6%	2		
20	Sweden	0.8%	1	Europe	4.8%
21	Ukraine	0.8%	1		
22	United Kingdom of Great Britain and Northern Ireland	3.2%	4		
23	Canada	2.4%	3	N. America	8.1%
24	United States of America	5.6%	7		
25	Brazil	1.6%	2	S. America	1.60%

*The highlighted countries are from the Middle East.

The response to question 4, "What is your area of expertise/study?" Is given in Table 5. The answer option "Other" received the following three responses: Human Resources, Information Science and Knowledge Management, and information and communications. These three responses can be added under the category Computer Science/Computer Engineering / IT, increasing the percentage of response to 60%. The education/teaching category received 14.4% followed by the sciences which received 7.2%. This shows that the main categories of respondents were from the IT, engineering and sciences sectors with very low responses from other fields. The main reason could be that the authors' contacts were mainly from IT and engineering; however, the survey clearly mentioned that people from all areas of expertise/study are invited to fill the survey, and contacts were urged to forward the survey link to their colleagues and friends.

The level of education of respondents is very high as it can be seen from Table 6 that 88% of the respondents have post-graduate education with 57.6% Doctorate and 30.4% Master degree holders. The rest of the respondents: 8% earned or are pursing their Bachelor degrees and 4% earned or are pursuing their High School diploma.

The respondents overwhelmingly consider that having access to technology is "very important", and this is very much expected given the technology era we are living in, Figure 4.

Table 4: Statistics on the age of respondents

Question 3: What is your age?		
Age category	Response Percent	Response Count
20 or younger	1.6%	2
21 – 30	14.3%	18
31 – 40	35.7%	45
41 – 50	28.6%	36
51 – 60	14.3%	18
61 – 70	4.8%	6
70 or older	0.8%	1

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 5: Area of expertise/study of survey participants.

Question 4: What is your area of expertise/study?		
Answer Options	**Response Percent**	**Response Count**
Business and Economics	3.2%	4
Computer Science/Computer Engineering / IT	57.6%	72
Education/Teaching	14.4%	18
Engineering	3.2%	4
Government, Law, Politics and Policy	0.8%	1
Language and Linguistics	4.0%	5
Medicine, Nursing and Health Sciences	5.6%	7
Philosophy, Ethics and Theology	0.8%	1
Sciences (Biology, Chemistry, Mathematics, Physics.)	7.2%	9
Visual and Performing Arts and Film Studies	0.8%	1
Other (please specify)	2.4%	3

Table 6: Level of Education

Question 5: What is the highest level of education you have completed?		
Answer Options	Response Percent	Response Count
In High School	3.2%	4
Graduated from high school	0.8%	1
2 year Diploma	0.0%	0
1st year of university	0.8%	1
2nd year of university	0.8%	1
3rd year of university	0.0%	0
4th year of university	0.0%	0
Earned a Bachelor's degree (4 or 5 years)	6.4%	8
Earned a Master degree	30.4%	38
Earned a Doctorate degree	57.6%	72

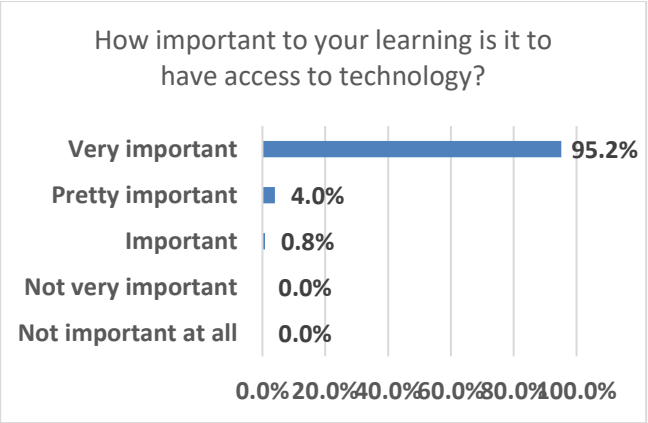


Figure 4: How important to have access to technology?

B. Application software technologies

This study investigated the capability of individuals with application software technology, Question 6, which most of us may be using on a daily basis, such as text processing, adding links to media files, creating and editing video and audio files, publishing websites, … etc.

The results show that all participants are capable of creating basic text documents with a word processor, ranging from advanced to acceptable levels. For creating text documents, using advanced features and adding multimedia links into documents almost 98% of the participants are capable of performing such tasks. Overall, more than 70% of the participants are capable of using all the application software technologies listed in Table 7. There are some technologies that not too many people were exposed to, such as creating and publishing a blog or online journal, programming a computer application, and designing and publishing websites.

Table 7: Statistics on how capable are individuals with application software technologies

Question 7: How capable are you with each of the following technologies to create or share documents?						
Answer Options	Very Capable Advanced	Capable Good	Somewhat Capable Acceptable	Not very Capable Poor	Never Used	Response Count
Creating text documents with a word processor	**79.70%**	**17.10%**	**3.30%**	**0.00%**	**0.00%**	123
Creating text documents using advanced features such as tables, images, formatting, macros... etc.	**70.20%**	**25.00%**	**3.20%**	**1.60%**	**0.00%**	124
Adding links to videos, audio, images into documents.	**63.70%**	**22.60%**	**11.30%**	**0.00%**	**2.40%**	124
Finding help from other people on the Internet using forms, message boards, social media ... etc.	**33.10%**	**37.90%**	**19.40%**	**4.80%**	**4.80%**	124
Programming a computer application or program for others to use	**32.00%**	**25.40%**	**18.00%**	**12.30%**	**12.30%**	122
Sharing or embedding video/audio files on websites.	**29.50%**	**36.10%**	**20.50%**	**4.90%**	**9.00%**	122
Creating and publishing a blog or online journal	**23.60%**	**24.40%**	**22.00%**	**11.40%**	**18.70%**	123
Creating and editing videos	**21.10%**	**29.30%**	**35.80%**	**7.30%**	**6.50%**	123
Creating and editing audio/sound files and recordings	**21.10%**	**30.90%**	**35.00%**	**5.70%**	**7.30%**	123
Designing websites using HTML, CSS, JavaScript	**17.90%**	**26.00%**	**26.00%**	**18.70%**	**11.40%**	123

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 8 provided the individual's usage of application software. What is interesting is that social media and word processing software are used daily by approximately 75% of the participants. Also, writing programs and using imaging software are used by approximately 15% of the participants. These software are specialized, and given the high level of education of these participants, this is expected. It is also shown that approximately 50% of the participants use these software at least once a month. On the other hand, playing computer games, using web development software, and video processing software are rarely or never used by most of the respondents with more than 60% of the respondents answered "rarely" or "never used".

Table 8: Statistics on how often individuals use application software

Question 8: How often do you use the following application software?						
Answer Options	Daily	Weekly/ Biweekly	Monthly	Rarely	Never	Response Count
Using Social Media websites (Facebook, Twitter, … etc)	75.80%	14.50%	3.20%	6.50%	0.00%	124
Word processing software (MS. Word, Apple iWork,... etc.)	74.80%	16.30%	4.10%	4.10%	0.80%	123
Writing programs using programming language (C++, Java, .etc.)	14.60%	14.60%	17.90%	30.10%	22.80%	123
Editing imaging software (paint, Photoshop, ... etc.)	14.40%	29.60%	22.40%	30.40%	3.20%	125
Playing computer games	8.30%	6.70%	15.80%	40.00%	29.20%	120
Web development software (HTML, CSS, Java Script, ... etc.)	4.90%	13.00%	21.10%	40.70%	20.30%	123
Video processing/editing software (Movie Maker, iMovie,etc.)	4.00%	9.70%	22.60%	50.80%	12.90%	124

The responses to the problems encountered when downloading application software, Table 9, show that all problems listed were encountered by at least 17% of the respondents with "the software is not free and I cannot afford buying it" receiving the highest percentage, 61.3%. Note that in, Question 9, the respondents can choose more than one answer, and those who have not encountered any problems may choose to skip the question; however, some of the respondents chose the answer option "Other" and wrote "nothing", "never used", or "no problems encountered". From the "other" responses, the respondents did not really state any problems except for one response which was "Not all software are fully localized for Middle East", and this is a real problem that many may face especially if they are not familiar with English or technical terminologies, which are very much needed during the installation process of any software.

Table 9: Problems encountered with application software

Question 9: Which of the following problems you have encountered with the application software in Question 8? (You may choose more than one answer).		
Answer Options	Response Percent	Response Count
Unable to download the software.	18.9%	20
Unable to install the software and decided to remove and not use.	17.0%	18
Had problems during installation and had to receive help online and/or from a friend.	17.0%	18
The software is not easy to use, the interface is not user-friendly.	24.5%	26
I always have to consult manuals, online help or YouTube videos (tutorials) for help.	32.1%	34
The software is not free and I cannot afford buying it.	61.3%	65
Other (please specify)	13.2%	14

C. OCR Software

In Question 10, it was asked if the individual used a scanner or have taken a photo/image of a text document which may or may not include text and images. The results show that 94.9% of the respondents have used a scanner. This was followed by Question 11, "Have you used OCR software to convert image files to editable text files?" It was interesting from the results to see that over 62% of the respondents used OCR software, either through their scanner OCR software or special OCR software. This question was followed by Question 12 on the languages used for OCR, which listed the top 10 spoken languages in the world, [42]. These are: Chinese, Spanish, English, Hindi, Arabic, Portuguese, Bengali, Russian, Japanese and Punjabi. An "other" option was also provided for the participants to include their language, if it is not on the list of languages. The "Other" option received 16 responses. The results show that the most used languages with OCR were English followed by Arabic then French, Table 10. Other languages with one or two responses include Spanish, Russian, Punjabi, Malay, and Swedish. French, Swedish and Malay were under the "Other" answer option where French received five responses, but Swedish and Malay one response each. In contrast with the rest of the responses, 9 responses of the "Other" option were "never", "not used", "nothing" … etc. Question 13 examined the usage of the top four best OCR software in 2017, and the results are given in Table 11.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 10: Responses for Question 12, "For which of the following languages have you used OCR"

Question 12: For which of the following languages have you used OCR software? (You may choose more than one answer).		
Answer Options	Response Percent	Response Count
Chinese	0.0%	0
Spanish	1.0%	1
English	83.3%	85
Hindi	0.0%	0
Arabic	44.1%	45
Portuguese	2.0%	2
Bengali	0.0%	0
Russian	2.0%	2
Japanese	0.0%	0
Punjabi	1.0%	1
French	4.9%	5
Malay	1.0%	1
Swedish	1.0%	1

Table 11 shows that the best OCR software, Adobe Acrobat, has been used by approximately 84% of respondents. On the other hand, all other software have slightly, rarely or never been used by most of the respondents. A minimum of 73% of respondents never used them. Also, it is observed that most of those who may have used these software rarely used them, and the only software with reasonable daily usage rate is Adobe Acrobat (30.7%) followed by ABBYY FineReader (1.9%). Similarly, in terms of competence in using the software, roughly the same rates for usage are reflected in the capability levels of individuals, Table 12. Overall, those who used the software daily show more competence in using the software. The survey investigated the problems which the users may have encountered during the installation process, Table 13. The results show that 41% of the respondents may need to use the software; however, they are surprised that the software is not free, and they cannot afford to buy it. In this study, 41% of the respondents consider this as a problem. In addition, 27.8% of the respondents faced difficulties, and they tend to consult technical support, manuals, tutorials, … etc., for help in order to resolve their issues. The "Other" answer option did not receive any responses which are worth mentioning, and most of the responses were "none", "never used", "Arabic OCR needs some work to reach the accuracy of English" … etc.

In addition, Question 16 presented the nine most available features in OCR software (given in Table 14) and asked the respondents if they know that such features are available in OCR software. Overall, the results showed that 56% of the users are familiar with most of the OCR available features.

Table 11: Statistics on the usage of OCR software

Question 13: How often have you used the following OCR software? and how often? (Note, Never, means you haven't heard about or used this software before?						
Answer Options	Daily	Weekly/ Biweekly	Monthly	Rarely	Never	Response Count
Adobe Acrobat	30.7%	14.0%	8.8%	29.8%	16.7%	114
OmniPage Standard	0.9%	0.9%	4.7%	19.6%	73.8%	107
ABBYY FineReader	1.9%	2.8%	0.9%	14.2%	80.2%	106
Readiris	1.0%	1.9%	2.9%	17.1%	77.1%	105

Table 12: Statistics on the level of capability of individuals with OCR software

Question 14: How capable are you with each of the following OCR software you have used?						
Answer Options	Very Capable Advanced	Capable Good	Somewhat Capable Acceptable	Not very Capable Poor	Never Used	Response Count
Adobe Acrobat	33.3%	30.6%	15.7%	6.5%	13.9%	108
OmniPage Standard	2.9%	8.6%	8.6%	8.6%	71.4%	105
ABBYY FineReader	2.9%	5.8%	2.9%	7.7%	80.8%	104
Readiris	2.0%	8.8%	5.9%	6.9%	76.5%	102

Table 13: Problems encountered with OCR software

Question 15: Which of the following problems have you encountered with the OCR software in Question 14? (You may choose more than one answer).		
Answer Options	Response Percent	Response Count
Unable to download the software.	13.3%	12
Unable to install the software and decided to uninstall and not use.	10.0%	9
Had problems during installation and had to receive help online and/or from a friend.	13.3%	12
The software is not easy to use, the interface is not user-friendly.	7.8%	7
I always have to consult manuals, online help or YouTube videos (tutorials) for help.	27.8%	25
The software is not free and I cannot afford buying it.	41.1%	37
Please enter your statement	22.2%	20

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Question 17 asks the respondents to rank the accuracy of the OCR software for the languages they have used. It was very surprising that over 91% of the respondents never used OCR for any language other than English and Arabic. Approximately 50% of the respondents ranks OCR for English as excellent, and 14.3% ranks Arabic as Excellent. However, 23.8% ranks OCR for English and 19.4% ranks OCR for Arabic as good. It is also noticed that more respondents used English OCR compared to any other language, this is followed by Arabic. The results may not be accurate and cannot be generalized since most of the respondents were from the Middle East, and very few people represented other languages mentioned in Table 15.

To further explore the use of OCR software, Question 18 asked if the respondents used OCR software to convert handwritten text documents into editable text documents. The results show that about a third of the respondents (32.4%) used OCR for handwritten documents. Also, Question 19 presents the OCR software accuracy level for handwritten documents.

Table 14: OCR Features

Question 16: Do you know that the following features are available with many OCR software?			
Answer Options	Yes	No	Response Count
Multi language support	72.2%	27.8%	108
The output file retains layout of original scanned paper document	63.2%	36.8%	106
The output file retains Fonts of original scanned paper document.	51.9%	48.1%	106
The output file retains Tables from original scanned paper document	52.8%	47.2%	106
OCR software can de-skew (rotate) image	50.9%	49.1%	106
Multi-page document recognition	58.1%	41.9%	105
Integrates with Cloud Storage.	41.0%	59.0%	105
Integrates with MS Office applications.	66.4%	33.6%	107
Integrates with HTML.	42.5%	57.5%	106
Average	56%	44%	106

Table 16, provides the results for Question 19, presenting the estimated accuracy from the user's perspective for the languages used by the respondents. The results show that a minimum of 93% of respondents never used OCR for handwritten documents. The accuracy level, as expected, is very low compared to printed documents. Here, English received the highest accuracy response rate of 18.8% as "Excellent" i.e., above 85% accuracy, followed by Arabic, 10.6%.

To conclude, the respondents were asked to provide the statement/s on how they feel about OCR. Table 17 shows the statement, "OCR is a very important technology and must be available as a feature in all text/image processing software," receives the highest response with 78.3%, followed by the statement, "OCR is very important to be used by universities and public libraries to convert old documents to digital documents," with 41.5%.

Table 15: Ranking the accuracy of OCR software for different languages using printed documents

Question 17: How do you rank the accuracy of OCR in the following languages? if applicable?						
Answer Option	Excellent (Above 85%)	Good (70% - 84.9%)	Acceptable (50% - 69.9%)	Poor (Below 50%)	Never Used	Response Count
Chinese	0.0%	3.6%	2.4%	1.2%	92.8%	83
Spanish	0.0%	3.8%	2.5%	0.0%	93.8%	80
English	49.5%	23.8%	9.5%	4.8%	12.4%	105
Hindi	1.2%	3.7%	2.4%	2.4%	90.2%	82
Arabic	14.3%	19.4%	16.3%	11.2%	38.8%	98
Portuguese	0.0%	6.0%	1.2%	0.0%	92.9%	84
Bengali	0.0%	1.2%	2.4%	2.4%	93.9%	82
Russian	0.0%	2.5%	4.9%	1.2%	91.4%	81
Japanese	0.0%	3.7%	1.2%	2.4%	92.7%	82
Punjabi	0.0%	1.2%	3.7%	2.5%	92.6%	81

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 16: Ranking the accuracy of OCR software for different languages using handwritten documents

Question 19: If your answer to Question 21 is yes, How was the accuracy? and in which language?						
Answer Option	Excellent (Above 85%)	Good (70%-4.9%)	Acceptable (50% - 69.9%)	Poor (Below 50%)	Never Used	Response Count
Chinese	0.0%	3.5%	0.0%	0.0%	96.5%	57
Spanish	0.0%	1.8%	0.0%	1.8%	96.5%	57
English	18.8%	17.4%	14.5%	4.3%	44.9%	69
Hindi	1.8%	0.0%	0.0%	5.4%	92.9%	56
Arabic	10.6%	9.1%	12.1%	6.1%	62.1%	66
Portuguese	0.0%	1.8%	0.0%	0.0%	98.2%	57
Bengali	1.8%	0.0%	0.0%	1.8%	96.5%	57
Russian	0.0%	1.8%	0.0%	1.8%	96.5%	57
Japanese	0.0%	1.8%	0.0%	1.8%	96.5%	57
Punjabi	0.0%	0.0%	1.8%	0.0%	98.2%	57

Table 17: Responses to "Select the statement that indicates how you feel about OCR"

Question 20: Please select the statement that indicates how you feel about OCR? (You may choose more than one statement).		
Answer Options	Response Percent	Response Count
OCR is a very important technology and must be available as a feature in all text/image processing software.	78.3%	83
OCR is a technology that is not mature enough to be used for some languages which written from right to left, such as Arabic.	21.7%	23
OCR is a technology that is not mature enough to be used for many languages.	16.0%	17
OCR is acceptable to be used for printed text; however, with handwritten text it is preferred not to be used.	19.8%	21
OCR is very important to be used in government offices to convert handwritten documents to editable files.	22.6%	24
OCR is very important to be used by universities and public libraries to convert old documents to digital documents.	41.5%	44
OCR is a saver for old documents and manuscripts from being lost/stolen/burned .... etc.	36.8%	39
Other (please specify)	8.5%	9

Finally, the respondents were given the chance to provide comments and/or suggestions about the survey. From the responses, almost all of the comments were "thank you", "nothing", and "good luck" comments except for the following three interesting comments: the first comment was "Thanks for providing me an opportunity to be part of your research and share my contribution. I think that expertise of university teacher's use of software and its related knowledge depends upon the availability of software in the institution. If an institution has enough sufficient software packages, then its next responsibility is to train its employees in using that software. So, institutions should have latest software and expertise in using and providing training in use of that software." The second comment was "I am not familiar with most of the software, and I think it is useless to have these many software". The third comment is, "I only use the software provided by the operating system, and those available on my computer at work." These comments summarize the reasons why people are not familiar with many software. We can say that we spend most of our time at work, and this influence the type of software we may use depending on our areas of expertise.

V. Limitations and Discussion

In general, document analysis software is very essential. Hence, OCR reduces time for processing data collection, which if done manually (data entry), takes longer time and is prone to human errors. The study findings may only benefit very limited number of people who are interested in English/Arabic OCR and image editing software and, therefore, not too many people participated in this survey. In addition, the low response rate could be because there are no monetary incentives for filling the survey. The results show that people were overwhelmed with the number of available software which most of them never heard about. Even though the Top four OCR software were listed, the results show that approximately 73% of the respondents never used or even heard about these software.

There are several limitations in the study findings. The study results cannot be generalized since the data collection was not purely random collected from faculty members in the IT sector to widen the sampling group to include individuals from all different sectors. The study reported approximately 58% of respondents are from the IT sector. The survey was distributed online to people who are familiar with using the Internet, sending emails, and using social media. The data collection was geographically distributed all over the world; however, 58% of the respondents were from the Middle East and about 71% of the Middle East respondents were from Saudi Arabia.

The accuracy rate reported by the commercial OCR software does not reflect the results obtained in this study. For example, the study reported the recognition rate for English, Latin script, as excellent with only 50% of the respondents, which is still not

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

100% accuracy as reported in [40]. Therefore, further investigation of the accuracy rates is needed for all language scripts including Latin scripts.

In the future, this study needs to be distributed to a wider range of people especially young people who are technology talented, and this survey did not reach them. In this study, 20% of the respondents were below 30 years old with only 1.6% were 20 or younger. Although this study provides insights into the familiarity of individuals with document and OCR technology corresponding to the age and gender, future research support is needed to understand the familiarity and capability of individuals with different software based on different age groups and gender.

The area of expertise and employment place of an individual pretty much have an influence on most of the types of software used in addition to the influence of the work environment which also influences the type of social media used. Another important finding from the survey is that technical support for the software in multi-languages, especially Arabic, is not available, which may be a factor that such software is not popular to use.

Currently, companies are providing free or limited usage of online OCR service or OCR mobile apps. This provides a real-time processing which can be used anytime from a personal computer or a mobile device. Future study should also evaluate this service and its possible replacement for current software installed on desktop computers.

A framework for testing OCR software is urgently needed to be able to test multilingual document analysis and OCR software. Databases of documents to test this software for all languages need to be developed in a similar fashion to be able to test different languages and provide comparable results for any language under any OCR platform. Therefore, the author recommends that more research is encouraged for different languages other than Latin, where the recognition rates are far below from those reported for Latin scripts, especially cursive scripts and printed handwritten scripts for all languages including Latin.

VI. Conclusion and Future Work

The OCR technology went through a major progress has progressed a great amount since it started in the early years of the last century. The study has confirmed that document analysis software and OCR, in particular, are very important and essential technology, and its use is seen in many applications around us. The recognition of cursive text in any language is still an active area of research. Higher recognition rates for printed or handwritten text is not possible without the use of contextual or grammatical information. The work carried out in this paper is the first of its kind and it sheds some light on the user perspective and their familiarity with document analysis software, especially OCR. The study concludes that a framework for testing OCR software is urgently needed to be able to test multilingual document analysis and OCR software. Finally, more research is encouraged for different languages other than Latin scripts.

Acknowledgments

The author would like to thank all colleagues and their friends who voluntarily filled the questionnaire survey for this study. In addition, the authors would like to thank the anonymous reviewers for their valuable comments and necessary suggestions that helped to clarify and improve the quality of the paper.

References

[1] B. B. Chaudhuri, (Ed.), “Digital document processing: major directions and recent advances,” *Springer Science & Business Media*, 2007.

[2] H. D. Wactlar, M. G. Christel, Y. Gong and A. G. Hauptmann, "Lessons learned from building a terabyte digital video library," *Computer*, vol. 32, no. 2, pp. 66-73, Feb, 1999.

[3] H. F. Herbert, "The history of OCR, optical character recognition." *Manchester Center*, VT: Recognition Technologies Users Association, 1982, ISBN 9780943072012.

[4] R. White, and T. Downs, “How computers work,” *Que Corp.*, 2007.

[5] S. Singh, "Optical character recognition techniques: a survey," *Journal of emerging Trends in Computing and information Sciences*, Vol. 4, no. 6, pp. 545-550, 2013.

[6] https://ocr.space/ocrapi, <Accessed 20 December 2016>.

[7] S. Naz, A.I. Umar, S. H. Shirazi, S. B. Ahmed, M. I. Razzak, and I. Siddiqi, "Segmentation techniques for recognition of Arabic-like scripts: A comprehensive survey," *Education and Information Technologies.* Vol. 21, no. 5, pp. 1225-1241, 2016.

[8] Y. M. Alginahi, “Computer analysis of composite documents with non-uniform background,” Thesis, University of Windsor, 2004.

[9] Y. M. Alginahi, D. Fekri, and M.A. Sid-Ahmed, “A neural-based page segmentation system,” *Journal of Circuits, Systems, and Computers*, 14(01), pp.109-122, 2005.

[10] Y. M. Alginahi, M. Mudassar, and M. N. Kabir, "An Arabic Script Recognition System." *KSII Transactions on Internet and Information Systems*, Vol. 9, no. 9, 3701-3720, 2015.

[11] I. Kaur and S. Mahajan, "Bilingual Script Identification of Printed Text Image," *International Journal of Engineering and Technology*, Vol. 2, no. 3, pp. 768-773, 2015.

[12] Y. M. Alginahi, “A survey on Arabic character segmentation,” *International Journal on Document Analysis and Recognition (IJDAR)*, Vol. 16, no. 2, pp.105-126. 2013.

[13] N. Fareen, M. A. Khan, and A. Durrani, "Survey of Urdu OCR: an offline approach," *Proceedings of the Conference on Language & Technology*, pp. 67-72, 2012.

[14] Parvez, Mohammad Tanvir, and Sabri A. Mahmoud. "Offline Arabic handwritten text recognition: a survey," *ACM Computing Surveys (CSUR)*, Vol. 45, no. 2, 2013.

[15] M. Shatnawi, Off-line handwritten Arabic character recognition: a survey. I*nternational conference on image processing, computer vision, and pattern recognition (IPCV)*, pp. 52 – 58, 2015.

[16] U. Saeed, "Automatic Recognition of Handwritten Arabic Text: A Survey." *Life Science Journal*, Vol. 11, no. 3s, 2014.

[17] D. N. Hakro, A. Z. Talib, Z. Bhatti, and G. N. Moja, “A Study of Sindhi Related and Arabic Script Adapted languages Recognition,” *Sindh University Research Journal-SURJ (Science Series)*, Vol. 46, no. 3, pp. 323-334, 2014.

[18] D. N. Hakro, A. Z. Talib, Z., and G. N. Moja, "Multilingual Text Image Database for OCR," *Sindh University Research Journal-SURJ (Science Series)*, Vol. 47, no. 1, 2016.

[19] D. N. Hakro, and A. Z. Talib, "Printed Text Image Database for Sindhi OCR," *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 15, no. 4, 2016.

[20] J. Sadri, M. R. Yeganehzad, and J. Saghi, "A novel comprehensive database for offline Persian handwriting recognition," *Pattern Recognition*, Vol. 60, pp. 378-393, 2016.

[21] M. Wahab, H. Amin, and F. Ahmed, "Shape analysis of Pashto script and creation of image database for OCR," *IEEE International Conference on Emerging Technologie*s (ICET 2009), pp. 287-290, 2009.

[22] M. Sagheer, C. He, N. Nobile, and C. Suen. "A new large Urdu database for off-line handwriting recognition," International Conference on Image Analysis and Processing, pp. 538-546. Springer Berlin Heidelberg, 2009.

[23] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri. "IFN/ENIT-database of handwritten Arabic words," *Francophone International Conference on writing and Document", CIFED'02*, Vol. 2, pp. 127-136, 2002.
[24] S. Mir, S. Zaman, and M. W. Anwar. "Printed Urdu Nastalique Script Recognition Using Analytical Approach." *IEEE 13th International Conference on Frontiers of Information Technology (FIT)*, pp. 334-340, 2015.
[25] R. Ahmad, M. Z. Afzal, S. F. Rashid, and S. Naz, "Semi-Automated Transcription Generation for Pashto Cursive Script," *J. Appl. Environ. Biol. Sci*, Vol. 6, no. 3S, pp. 96-101, 2016.
[26] S. Nasrollahi, and A. Ebrahimi, "Printed Persian Subword Recognition Using Wavelet Packet Descriptors," *Journal of Engineering*, 2013.
[27] D. V. Kaur, "Comparison Analysis of Post-Processing Method for Punjabi Font." *International Research Journal of Engineering and Technology (IRJET)*, Vol. 4, no. 1, pp. 1206 – 1211, January 2017.
[28] Q. Chen, B. S. Yuan, X. LI, H. Y. REN, and J. H. Zhang, "Printed Uyghur Character Recognition Based on Template Matching," *Computer Technology and Development*, Vol. 4, 2012.
[29] A. A. Sanjrani, J. Junaid, M. Bakhtyar, W. Noor, and M. Khalid, "Handwritten Optical Character Recognition system for Sindhi numerals," *International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, pp. 262-267, 2016.
[30] Z. Jan, M. Shabir, M. A. Khan, A. Ali, and M. Muzammal, "Online Urdu Handwriting Recognition System Using Geometric Invariant Features," *Nucleus*, Vol. 53, no. 2, pp. 89-98, 2016.
[31] Ahmad, Riaz, Muhammad Zeshan Afzal, Sheikh Faisal Rashid, Marcus Liwicki, Andreas Dengel, and Thomas Breuel. "Recognizable units in Pashto language for OCR." *13th International Conference on Document Analysis and Recognition (ICDAR'15)*, pp. 1246-1250, 2015.
[32] M. Kazemi, M. Yousefnezhad, and S. Nourian. "A New Approach in Persian Handwritten Letters Recognition Using Error Correcting Output Coding," *Journal of Advances in Computer Research*, Vol. 6, no. 4, pp. 107-124, 2015.
[33] G. Z. Campus, G. N. NIET, and P. Bathinda, "Review: A Literature Survey on Text Segmentation in Handwritten Punjabi Documents," *Research Cell: An International Journal of Engineering Sciences*, Vol. 3, pp. 174 – 178, December 2014.
[34] H.Y. Ren, B.S. Yuan, and Y. Tian, "On-line & Handwritten Uyghur Character Recognition Based on BP Neural Network," *Journal of Microelectronics & Computer*, Vol. 8, p.060, 2010.
[35] D. N. Hakro, I. A. Ismaili, A. Z. Talib, Z. Bhatti, G. N. Mojai, "Issues and challenges in Sindhi OCR," *Sindh University Research Journal-SURJ (Science Series)* Vol. 46, no. 2, 2014.
[36] E. Borovikov and I. Zavorin. "A multi-stage approach to Arabic document analysis. Guide to OCR for Arabic scripts," 2012, *Springer*, London, pp. 55-78.
[37] R. Holley, "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitization programs." *D-Lib Magazine*, Vol. 15, no. ¾. 2009.
[38] P. R. Dussawar and P. B. Jha. "Text Extraction from Complex Color Images Using Optical Character Recognition," *International Journal of Science and Research (IJSR)*, Vol. 4, no. 7, pp. 730 – 735, July 2015.
[39] C. Y. Suen, R. Plamondon, A. Tappert, A. J. W. M. Thomassen, J. R. Ward, and K. Yamamoto. "Future challenges in handwriting and computer applications." *Third International Symposium on Handwriting and Computer Applications, Montreal*. May 1987.
[40] http://www.toptenreviews.com/business/software/best-ocr-software/, <Accessed 12 November, 2016 >
[41] http://download.cnet.com/s/ocr/, <Accessed 15 November, 2016 >
[42] https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world, <Accessed 29 November, 2016 >

Yasser M. Alginahi, earned a Ph.D., in Electrical Engineering from the University of Windsor, Ontario, Canada, and a Master of Science in Electrical Engineering from Wright State University, Ohio, U.S.A. He is a licensed Professional Engineer, Ontario, Canada, a member of Professional Engineers Ontario, a senior member of IEEE and IACSIT since 2010. Dr. Alginahi is an Associate Professor at the Department of Computer Science, Deanship of Academic Services, Taibah University. He is also the consultation unit coordinator at the IT Research Center for the Holy Quran and its Sciences, Taibah University. Dr. Alginahi is on editorial board of several international journals. He published a book entitled "Document Image Analysis" and he has published over 80 journal and conference papers. He worked as Principle Investigator and co-Principle Investigator on many funded research projects by the Deanship of Scientific Research at Taibah University and other organizations such as King Abdul-Aziz City of Science and Technology. His current research interests are Information Security, Document Image Analysis, Pattern Recognition, OCR, Modeling and Simulation, Intelligent Transportation Systems and Numerical Computations.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Digital Holy Quran Authentication, Certification and Standardization: A Survey

Yasser M. Alginahi[1,2], Muhammad Khurram Khan[3], Omar Tayan[2, 4], Mohammed Zakariah[5],
[1]Dept. of Computer Science, Deanship of Academic Services, Taibah University, KSA
[2]IT Research Center for the Holy Quran and Its Sciences (NOOR), Taibah University, KSA
[3]Center of Excellence in Information Assurance (CoEIA), King Saud University, KSA
[4]College of Computer Science and Engineering, Dept. of Computer Engineering, Taibah University, KSA
[5]Research Center of College of Computer and Information Sciences, King Saud University, KSA
yasser.m.alginahi@ieee.org, mkhurram@ksu.edu.sa , otayan@taibahu.edu.sa, mzakariah@ksu.edu.sa

***Abstract*- The digital authentication, certification and standardization of Information Technology (IT) applications of the Holy Quran has attracted researchers and organizations to explore its importance and requirements. This is an emerging area of research that needs more attention since it investigates sensitive digital documents and applications. The objective of this study is to gather information from users, developers, designers, decision makers and organizations on their perception on authentication, certification and standardization of IT applications for the Holy Quran. This study is based on a questionnaire survey which was distributed online to highly education individuals from all over the world. The study emphasized the need for more security and quality assurance in IT Quran applications. In addition, standardization and certification through authentic Quran organizations are encouraged to validate the Quranic content. The study results and analysis were extracted from over 80 questions divided between six categories of participants. The survey received 500 responses from 31 countries and five continents. The findings of the study are considered a step towards the importance of secure sensitive documents in all different areas of research particularly digital Quran content.**

***Keywords*- Quran, Authentication, Security, Quality of Service, Standards**

I. INTRODUCTION

Quran is the sacred book of Islam which was revealed to Prophet Mohammed (peace be upon him) over 14 centuries ago. Quran is usually read from paperback format called Mushaf; however, nowadays the recent advancements in technology allowed the use of smart digital gadgets which are portable and adopted by almost everyone [1]. Therefore, the rapid advancements in technology with the worldwide spread of the Internet allowed for the dissemination of online digital multimedia content. The increase in speed and storage capabilities allowed for a rapid increase in the digitization of all sorts of multimedia publications especially the digitization of printed documents has seen an exponential jump as archives are digitized and offices are eliminating papers and becoming paper free. Therefore, Quranic organizations and research centers exploited this movement of digitization to benefit in the development of Quranic Web and mobile apps. In this section, the most recent research work related to Quan computing i.e., security, authentication and standardization is presented.

Due to the sensitivity of Holy Quran, it is very crucial to authenticate the verses or parts of the Quran available through webpages to avoid intentional and unintentional distortion/forgery to Quranic verses. The goal is to use information retrieval techniques to check for documents that include unauthentic Quranic

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

verses, therefore, authentication techniques are required to check the world-wide web for any unauthentic Quranic verses. Thus, this is possible but false results may appear if other different Quranic readings text which may be found in some websites [8].

The work by Tayan and Alginahi, conducted a survey to investigate the significance of online and offline software applications and their benefits and effectiveness for users. The study showed that approximately 80% of participants had used technology in Quran memorization, technology is more appealing to the younger generation and 50% of the participants preferred Internet/software or handheld/portable technologies as compared with traditional methods. The study concluded with several challenges from the user-perspective these are: English-language barrier since many applications are in English, Technology illiteracy, limited Internet resources and connectivity. In addition, the participants provided suggestions for new applications and modifications to existing technology [2]. In another work, Tayan and Alginahi presented a review of different multimedia watermarking techniques which are applicable to sensitive digital content and discusses how those approaches can accomplish the required protection [3].

The work in [4] presented a review on the types of Information Security (IS) aspects which include; data-storage, in-transit data and data access prevention for unauthorized users. The work presented the main challenges in IS with regards to vulnerabilities and data-breaches in addition to discussing the mechanisms for enhancing data protection, IS policies and standards for protecting data content [4].

The survey in [5] provided a review on recent researches on Digital Holy Quran authentication, protection and integrity authenticity, it focused on analyzing and categorizing the existing research related to preserving and verifying the content integrity of the Holy Quran. It concluded by a recommendation to develop a reliable universal database of authentic and verified Digital Quran and hadith content, in addition, to developing a Real-Time Quran Verse Detection Expert System with improved accuracy and precision [5].

The research paper “Digital Quran Computing: Review, Classification, and Trend Analysis” by Zakariah et al. provided a subject/theme categorization of digital Quran research based on the topical trends then a discussion of their key features, limitations, and research directions. Next, a set of recommendations regarding security and authentication, standardization and quality of service, unified translation, and E-Learning Approaches and Quran Knowledge base were presented. Finally, the authors concluded by discussing the open challenges in digital Quran Computing [6].

In [7], Sabbah and Selamat proposed a framework authenticity detection method for text Quranic verses extracted from online forum posts based on computing numerical Identifiers of words in the detected text then comparing these identifiers with Identifiers of original Quranic manuscript. The results show that the average accuracy rate was 62%, precision 75% and recall 78%. The authors stated that their future work will incorporate computational intelligence methods to increase the authentication which will also involve other Quran media such as sound, images and video to improve the detection.

Alsmadi and Zarour designed a tool to evaluate the integrity of the wording in the e-versions of the Quran by generating the metadata related to all words in the Quran i.e., to preserve the counts and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

locations. This is like hash algorithms used in checking the integrity of data files in a disk, as a result a tiny modification will result in a different hash function. The results show that hashing verification can be a good candidate for the automatic Quran authentication process with high confidence [8].

The exponential increase in the number of websites on the Internet the vulnerability of sensitive information, such as digital Quran, maybe at stake. From the huge number of Islamic websites not many of them present digital copies of the Quan that are verified and certified by an authentic source. The work by Mostafa and Ibrahim propose a computer-based system on the public key infrastructure and the digital signature to secure and verify the content of Holy Quran script on the web [9]. The authors claim its applicability to the application of the Holy Quran.

Digital copies of fake or distorted Quran have been detected online and in some Quran mobile apps. Therefore, to eliminate such forgery, by authentication, which may be intentional or unintentional will solve most the threats and provide confidence in users to use such web and mobile applications. The study in [10] presents a comprehensive research survey of works conducted in the area of Quran authentication from an information security perspective. This study concluded the urgent need to provide a good content security, and integrity of digital Quran [10].

Cryptography have been used to secure data against tampering attempts and protects the data integrity. The work by AlAhmad, Alshaikhli, and Jumaah proposed a cryptography algorithm (Combination between AES and RSA Cryptography Algorithms (CARCA)). The results show that the CARCA method, with the two encryption algorithms, showed a boost and improvement in the protection of the Digital Holy Quran Hash Digest [11].

The works in [12 - 13] presented a watermarking method to enable the authentication and detection of the image forgery on the digital Quran images. The proposed method uses two layers of embedding scheme. First, the discrete wavelet transforms are applied to decompose the host image into wavelet prior to embedding the watermark in the wavelet domain. Then, the watermarked wavelet coefficient is inverted back to the spatial domain where the least significant bits are utilized to hide another watermark. Next, a chaotic map is used to blur the watermark to secure it against the local attack. This technique provides high watermark payloads, though preserving the image quality.

On the issue of certification, Khan, Siddiqui and Tayan presented a digital Quran certification framework by utilizing modern digital authentication and certification techniques i.e., a certification authority and religious scholars follow a rigorous procedure that checks the requirements process for certification and upon approval a digital certificate is issued for the application. This framework controls the digital content of the Holy Quran and doesn't allow any modifications by the users [14].

To the best knowledge of the authors the literature provides a good quality of security and authentication, however, very scarce work on Quran certification and no work available on standardization of digital Quran. As presented in this section most of the available Quran authentication techniques have been used for the purpose of research and no commercial implementation of such techniques is provided. On the issue of certification digital copies of the Quran what is available is a framework for certification. On the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

other hand, the issue of standardization has not been address in literature. Therefore, the available research work is considered a positive step for researchers in this area to come together and set procedures and standards for security and authentication, certification and standardization of digital Quran.

In this paper, the security and authentication, certification, and standardization challenges regarding the development of Quran apps are presented from the perspective of users, developers, designers, decision makers and institutions. The study was based on a survey conducted to gather information on technology adoption for reading the Quran, security and quality control issues in development of Quan software (mobile and web applications), in addition to standardization issues of institutions and software. This work is unique as it gathered feedback/perspective from six different categories to identify all aspects of authentication, certification and standardization from the perspective of the users, developers and programmers, designers, managers, organizations and publishers.

II. Methodology

A semi-structured questionnaire survey was developed using monkeysurvey.ca and distributed online to the authors contacts who are mainly highly educated individuals from all over the world. The survey was conducted to discover This survey is delivered and distributed using a range of social media tools, and its findings are taken to characterize sampling of English/Arabic speaking users of Quran products. The main objective from this study is gathering the perspective/feedback of different categories of individuals and organization on the requirements and importance of digital authentication, Certification and Standardization of IT applications for the Holy Quran. Before the survey was conducted it was piloted to ten subjects who were interviewed so that to provide their feedback on the questions of the survey and comment on its clarity. Next, the semi-structured survey was designed then distributed online to over 3500 emails and social media accounts including WhatsApp. The response rate of the survey was approximately 14%. The survey was distributed in both English and Arabic. The total number of respondents was 500, from whom 307 filled the English survey and 193 filled the Arabic survey. The survey was designed using monkeysurvey.ca, the questions types include, Yes/No, Multiple choice Questions, and open-ended questions.

TABLE 1
THE PERSONAL PROFILE QUESTIONS ASKED

No.	Question	Type of Question	Completed Responses		Skipped Responses	
			Arabic	English	Arabic	English
1	What is your gender?	Multiple Choice	184	300	9	7
2	Which age group do you belong to?	Multiple Choice	184	302	9	5
3	What is your highest qualification?	Multiple Choice	184	301	9	6
4	Your employment status?	Multiple Choice	181	302	12	5
5	What is your Nationality?	Dropdown	173	292	20	15
6	Which category/occupation do you belong to?	Multiple Choice	185	289	8	18

The survey is divided into two parts, the first part includes personal questions and the second part includes specific questions targeting certain categories of individuals and organizations by requesting them to provide their feedback on questions related to requirements and importance of digital authentication,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

certification, standardization of IT applications for the Holy Quran. The survey with questions on the personal profile of the respondents, including, age, gender, nationality, employment status and education level. Table 2, shows that the highest response rate for this survey was from the age group 35 – 44 years old with approximately 41%, followed by 28% for the age group 45 – 54, then 19.8% for the age group 45 – 55 and 8.2% for the age group 55 - 64. It is noticed that the age groups below 25 and above 65 counted for about 3% of the respondents.

TABLE 2
AGE GROUPS OF PARTICIPANTS

Answer Options	Response Percent	Response Count
15-24	2.1%	10
25-34	19.8%	96
35-44	40.9%	199
45-54	28.0%	136
55-64	8.2%	40
65+	1.0%	5

From both surveys (Arabic/English) conducted, it was noticed that 44.6% of the participants are females and 55.4% are males. Table 3 shows that the minimum completed education level was a high school diploma and approximately 89.3% of the participants have a bachelor or graduate degree with approximately 51% hold a Doctorate degree which shows that most of the participants are highly educated.

TABLE 3
THE EDUCATION LEVEL OF PARTICIPANTS

Answer Options	Response Percent	Response Count
High School	2.9%	14
Associate/Vocational/Trade Degree	0.2%	1
Certified Professional/Diploma Holder	0.2%	1
Bachelor Degree Holder	5.2%	25
Master Degree Holder	26.2%	127
Doctorate Degree Holder	50.9%	247
Postdoctoral	12.2%	59
Prefer not to say	2.3%	11

The employment status of the participants shows that 91% are employed full-time, part-time or self-employed, 6% are students, 0.4% are retired, 1.4% unemployed and 1.2% preferred not to say. The nationalities of the participants included 31 countries from five continents. The following are the countries of the participants: United States (3), Algeria (38), Australia (4), Bahrain (3), Bangladesh (2), Canada (7), China (1), Egypt (34), Ethiopia (1), India (33), Indonesia (4), Iran (3), Iraq (7), Jordan (65), Kuwait (6), Lebanon (3), Libya (4), Malaysia (78), Morocco (9), Nigeria (3), Oman (7), Pakistan (54), Palestine (10), Saudi Arabia (23), Sudan (18), Spain (1), Syria (10), Tunisia (4), Turkey (3), United Kingdom (10), and Yemen (17). The number between parenthesis indicates the number of participants from each country.

After responding to the personal profile section of the survey, the participants have to choose the category they belong to from the six categories shown in Table 4 before moving to the second part of the survey and answer the corresponding questions related to that category. Table 4, provides the categories and the number of respondents in each category.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 4
THE CATEGORIES OF PARTICIPANTS

Name of Category	Response Percent	Response Count
1- A general IT/Software User or Advanced User	47.3%	224
2- Software/Mobile/Digital Application Developer	18.4%	87
3- Graphics Designer	1.9%	9
4- Organization Manager/Director/Decision Makers/Owners/CEOs	22.8%	108
5- Islamic Institutes	6.8%	32
6- Holy Quran Publishers	3.0%	14

Even though Table 4 shows that 95% of the 500 participants completed the personal profile by choosing a category only 350 participants, i.e., 70% of the total no. of participants, completed the questions of the category they chose in Table 4. Finally, Table 5 provides the responses for each category. The results show that 51.4% are general IT/software users, 18.3% are software developers, 22.3 managers and decision makers, with less than 7% Islamic institutes and Holy Quran publishers.

TABLE 5
THE RESPONSE RATE OF PARTICIPANTS FOR EACH CATEGORY

Name of Category	Response Percent	Response Count
1- A general IT/Software User or Advanced User (15 Questions)	51.4%	180
2- Software/Mobile/Digital Application Developer (16 Questions)	18.3%	64
3- Graphics Designer (8 Questions)	1.1%	4
4- Organization Manager/Director/Decision Makers/Owners/CEOs (15 Questions)	22.3%	78
5- Islamic Institutes (9 Questions)	5.7%	20
6- Holy Quran Publishers (11 Questions)	1.1%	4

III. Results Analysis & Discussion

This section presents the analysis and discussion for the results from the six survey categories.

A. General IT/Software User or Advanced User

The list of questions for this category is provided in Table 6.

TABLE 6
THE GENERAL IT/SOFTWARE USER OR ADVANCED USER CATEGORY QUESTIONS

No.	Question	Response Count
1	How often do you recite Qur'an?	180
2	Do you also recite Qur'an on-line?	180
3	What is the reason for not reciting the Qur'an on-line? (if the previous answer is NO)	69
4	Is your level of confidence in digital Quran authenticity equal to when compared with printed Quran books?	180
5	Do you use Digital Devices to read Qur'an (Smartphones/Tablets/ digital diary etc.)?	181
6	Why you do not recite the Qur'an on a digital device? (if the above is NO)*	47
7	In your opinion, are the current digital copies of the Quran available on different digital devices authentic?	179
8	In your case, do you prefer a digitally signed and 100% authentic copy of the Quran on-line or on digital device?	180
9	Did you ever encounter a fake copy of Qur'an available on-line or on a digital device?* Please provide URL.	177
10	What are the main facilities/services you use from online, web-based or smartphone Quran apps?*	186
11	Do you think it is necessary for a Quran Authentication Body to monitor and endorse the digital copies of the Qur'an Worldwide?	181
12	Are you pleased with the features of most online/smartphone Quran apps?	176
13	Please state the name of your favorite Quran Application smartphone or URL:	72
14	Are you pleased with the quality of information provided by most apps?	175
15	What is your vision/comments/recommendations of how new technologies could be used to develop and improve Quranic/Related Smartphone Apps, Web Apps or user interaction further?	79

*Question has an "Other" option or "comment box"

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The results show the users response to questions related to their personal use and confidence in digital Quran, in addition to their perception on authenticity. Table 7 provides the response of the participants to the question "How often do you recite Quran?" It shows that over 65% recite the Quran daily or on alternative days and 14.4% recite it weekly which means that at least 80% of respondents reads Quran at least once a week. Also, 62.2% recite Quran online besides reading from a paper copy of the Quran, and 37.8% only prefer reading from paper copies of the Quran. From those who do not read Quran online, 92.8% feel uncomfortable reading from a computer or smart mobile screen and 7.2% specified that they are not Internet users. However, the response to "Do you use digital devices (Smartphones/Tablets/Digital diary …etc.) to read Quran?" shows that 82.3% use digital devices compared to 17.7% who do not use any gadgets to read Quran. The reasons for not reading from digital devices include: medical reasons, feeling uncomfortable, feeling uncertainty of the authenticity of the Quran, prefer to listen from Musha'af (paper copy of the Quran), and getting more benefits from reading from Musha'af. These results emphasize the importance of reading the Quran regularly and the need to provide ways to make it easier to recite and memorize it.

TABLE 7
RESPONSE OF THE PARTICIPANTS TO THE QUESTION "HOW OFTEN DO YOU RECITE QURAN?"

Answer Options	Response Percent	Response Count
I have memorized the Qur'an by heart	2.2%	4
Daily	46.7%	84
Alternate days	16.7%	30
Once/Week	14.4%	26
Once/Month	4.4%	8
Once/Year	1.7%	3
Rarely	5.6%	10
Prefer not to say	8.3%	15
Doesn't apply to me	0.0%	0

The level of confidence in digital Quran authenticity compared with printed Quran paper back could be another reason why people do not use or trust digital copies of the Quran. The results show that 41.6% of the respondents never had any peculiar feelings at all in regard to the level of confidence in digital Quran authenticity, 20.6% always feel that the contents may be produced by an un-authentic source, 17.8% always have a feeling that the contents may be forged/modified or not properly scrutinized and 21.1% never thought about it.

Responding to the question "In your opinion, are the current digital copies of the Quran available on different digital devices authentic?" the response shows that 21.2% believe that the digital copies of Quran are authentic and 7.3% think otherwise. However, 60.3% are not sure and 11.2% never thought about it. This shows that many people assume what they download from app stores is authentic and such an issue does not cross their minds. In addition, most of the people do not memorize the Quran and may not even notice if there is a mistake. Therefore, the next question "do you prefer a digitally signed and 100% authentic copy of the Quran on-line or on digital device?" which addresses authenticity shows that 85% think it is very essential and will give authenticity a very high preference, 12.8% agree and may prefer 100% authenticity and 2.2% do not prefer digitally signed copy of the Quran. In response to the Question

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

"Did you ever encounter a fake copy of Qur'an available on-line or on a digital device?" 12.4% mentioned that they encountered a fake copy of the Quran. The True Furqan is one main example provided by participants, some asked to refer to google images and others could not recall the names of the websites that they encountered a copy of fake or modified Quran. This raises the question "Is it necessary for a Quran Authentication Body to monitor and endorse the digital copies of the Qur'an Worldwide?" and as expected the response of the participants highly emphasize this with 98.9% giving it a high preference.

The use of Internet or digital devices to recite the Quran also helps people use other facilities or services that are available within online Quranic/Islamic websites and/or digital apps. Table 8 shows that besides reciting the Quran, reading the Tafseer (Quran explanation) and using search tools are the most used services. Besides the services provided in Table 8, other services include learning Quranic Arabic, reciting Zikr (supplications), listening to YouTube, finding prayer times and Qibla direction, preparing Islamic lectures and sharing Quranic services on social media. The response count is more than the number of participants since this question allows for multiple answers.

TABLE 8
THE MAIN FACILITIES/SERVICES THE PARTICIPANTS USE FROM WEB AND MOBILE APPLICATIONS

Answer Options	Response Percent	Response Count
Quran text & recitation only	27.3%	143
Tafseer	23.1%	121
Use of those apps to help memorize Quran	7.6%	40
Le'rab (Grammar) of the Holy Quran.	5.0%	26
Search Facility	19.5%	102
Phonetic Search	5.7%	30
Statistical Analysis	4.8%	25
Retrieving, Printing contents or resources.	5.9%	31
Other (Please specify) ----------	1.1%	6

The survey also shows that 70.5% of participants are pleased with the features of most online/smartphones Quan apps and 73% are pleased with the quality of information provided by most apps. Therefore, there are many Quran apps and websites which the participants favor to use, however, the mostly mentioned are: Ayat (provided by King Saud University http://quran.ksu.edu.sa/), iQuran Lite (An IOS app), Quran Android, Quran Explorer http://www.quranexplorer.com/quran/, and Tanzil (http://tanzil.net/). To conclude this category the participants were asked to provide their comments/suggestions/vision on how new technologies could be used to develop and improve Quran related smart apps and web applications. Finally, the participants provided their comments and recommendations on how new technologies can be used to develop and improve Quranic related smartphone and web apps. The participants provided many comments, suggestions and recommendations. Some of the design features the participants would like to see in Quran apps are: zooming, contextual search, adjustable font size, high sound quality, multi-language, Tajweed learning, contact a scholar, daily random lessons and stories, pop up Word by word meaning while reciting Quran, Daily alarm/reminder, quiz, game-based feature to keep motivation, statistical analysis, advanced research and navigation features. Other features or apps the participants would like to see for or in Quran apps include, interactive and user-friendliness, offline accessibility, free of charge, robustness, ad free, free of charge, encapsulates

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

different features in one app such as qibla direction, prayer time … etc., user friendly voice-driven apps, an app that could match reciter's voice with the database then give rating of the quality of the recitation, features for people with special needs. Regarding security and authentication, the participants emphasized that authentication should be ensured as it increases the confidence of the user with secured apps. In addition, the participants provided the following comments and suggestions:

- Develop one global source of authentic digital Quran provider.
- Use cryptographic algorithms to ensure data integrity.
- Develop a security tool that is compatible with most Quran apps to be used to check originality of the Quran.
- Develop a secure cloud-based repository for Quran related resources and make it available as a web service for all mobile and web apps. Any Quran related resources must be first authentic before being used in designing apps.
- One or more Islamic authentication companies must check all available digital apps and issue them tags for authentication or rejection, which will make the users aware of authentic apps.
- Provide in app capability for real time authentication.
- Design a plug in that checks Quran verses online.
- Develop an agent program to keep monitoring the Quran text for correctness. Such text may be any uploads related to Quran with the help of popular social media and other websites that may contain such contents.

B. Software/SmartPhone/Digital Application Developer

The list of questions for this category is provided in Table 9. The results provide the users response to questions related to their personal experience as web and mobile application developers.

TABLE 9
THE "SOFTWARE/SMARTPHONE/DIGITAL APPLICATION DEVELOPER" CATEGORY QUESTIONS

No.	Question	Response Count
1	Are you involved in Digital Application Development for Smart devices or Web-Apps?	64
2	Did you ever develop any online Quran application that involved reading or reciting Quran using a Smart device?	64
3	Do you think the digital Quran which you have used in your application is 100% authentic and is free of any intentional/unintentional errors?	63
4	Do you consider "Content or Quranic Script Integrity/Authenticity" as a Primary Measure while developing a Quran application?	63
5	Do you apply Modern Image Processing or Signal processing techniques to prevent any tampering attempts?	62
6	What type of Signal-Processing or Image-Processing tools do you use?*	53
7	Do you follow any particular information security standards while developing applications?*	60
8	Which Security Standards do you follow? (If the previous answer is YES)*	26
9	Do you consider Secure Coding while developing such applications?	60
10	Did you ever witness any case in which the integrity/security of a Digital Quran application was compromised due to unsecure coding or development?*	61
11	Do you follow Quality Assurance standards while developing such applications?	59
12	Which Quality Standards do you follow? (If answer for the above question is YES)*	27
13	Will you consider to apply Security Standards, Techniques and Measures to prevent any type of forgeries while developing Quran Applications?	60
14	Do you think it is necessary for a Quran Authentication Body to monitor and endorse the digital copies of the Qur'an Worldwide?	61
15	Will you prefer validating your Quranic resources through a Quran Authentication Body?	61
16	What is your vision/comments/recommendations of how new technologies could be used to develop and improve Quranic/Related Smartphone Apps, Web Apps or user interaction further? Please specify:	26

*Question has an "Other" option or "comment box"

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The results of the survey show that 73.4% of the participants are involved in Digital Application Development for Smart devices or Web-Apps and 28.1% developed online Quran application that involved reading or reciting Quran using a Smart device. Approximately One third of the participants think the digital Quran which they have used in their applications is 100% authentic and is free of any intentional/unintentional errors, and 7.9% use no verification procedures during apps development. On the other hand, 42.9% are "Not sure" and 17.5% "Never thought about it". However, 90.5% of the participants consider Content or Quranic Script Integrity/Authenticity as a Primary Measure while developing a Quran application.

Modern image processing is one of the main technologies that can be used in designing applications and only 22.6% of the participants have used these tools to prevent any tampering attempts. The response rate to this question is very low given that fact that 51.6% responded with "No" and 25.5% "Never thought about this issue". This question is answered by developers who may or may not been have designed Quran applications, and therefore the response rate may not be reflective of the actual developers who designed Quran apps before. The image processing tools used by the participants include MATLAB image processing Tool-Box (30.2%), open source tools (41.5%), proprietary tools such as Adobe (9.4%) and in house dedicated tools (18.9%).

In response to the question "Do you follow any particular information security standards while developing applications?" 28.3 of the participants follow IS standards, 56.7% do not follow any standards, and 15% "Never thought about using any standards". The list of standards used is provided in Table 10. The response to the "Other" option includes. mathematical creteria, HCI standards, Nist, rfc, dedicated in house standards. Following this, the survey asked if the developers use secure coding while developing such applications, 46.7% responded with "Yes", 23.3% responded with "No" and 30% responded with "I never thought about secure coding.

TABLE 10
THE LIST OF SECURITY STANDARDS USED BY PARTICIPANTS

Answer Options	Response Percent	Response Count
ISO 9001	50.0%	13
ISO 27001	7.7%	2
PCI-DSS	0.0%	0
FFIEC	0.0%	0
Other (please specify)	42.3%	11

The security of the digital Quran should not be compromised when designing Quran applications. Therefore, security and quality assurance standards are very important in designing software. The results show that only 4.9% of the participants have witnessed some digital Quranic apps with modified/forged due to unsecure coding or development and in response to "Do you follow Quality Assurance standards while developing such applications?" 35.6% say "Yes", 40% say "No" and 23.7 say "I don't Know". The quality standards followed by the participants are given in Table 11. The "Other" option include, mathematical creteria, dedicated, and I use the Quranic text that is known to have been validated fromTanzil.org. Also, 86.7% of the participants expressed their interest to consider applying Security

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Standards, Techniques and Measures to prevent any type of forgeries while developing Quran Applications. Also, 98.4% think it is necessary for a Quran authentication body to monitor and endorse the digital copies of the Qur'an worldwide and 95.1% prefer validating Quranic resources through a Quran Authentication Body.

TABLE 11
THE LIST OF QUALITY STANDARDS USED BY PARTICIPANTS

Answer Options	Response Percent	Response Count
ISO 9004	51.9%	14
ISO 19011	3.7%	1
AS9100	0.0%	0
MBNQA	3.7%	1
Lean/Six Sigma/TQM	3.7%	1
McCall Quality Factors	0.0%	0
Other (please specify)	37.0%	10

Finally, the participants provided their vision, comments and recommendations of how new technologies could be used to develop and improve Quranic/Related Smartphone Apps, Web Apps or user interaction. Some of the participants comments and recommendations are:

- Develop interactive apps.
- Provide technical support for monitoring and addressing concerns from users.
- Give more attention to information security and quality assurance.
- Provide the Quran text as a service for developers in a way that they can only integrate it without having to modify.
- Watermarking can be used to tackle the security issue.
- Protect Quran applications by using information security algorithms.
- Follow very rigorous information scrutiny and authenticity procedure for the content by performing intrusion testing to make the application secure.
- Follow well known and secure programming methodologies.
- Develop an intelligent Quran content checker.
- Quranic Text and Translations must be verifiable by the user of an app.
- Anti-tampering feature with the originality of Quran should be included in such application to alert the users.
- Regulatory bodies (in collaboration with scholars) should be formed to monitor the development and distribution of Quran app and empowered in order to take action in case of a tampering case is detected.
- Authentic Digital Quranic text should be made available to Quran digital developers. It should be in form that is flexible enough to be used in all platform.
- Using mobile solution networks the authentic Quranic content can be disseminated among the social users of the communities.

C. Graphic Designer

The list of questions for the Graphic Designer category is provided in Table 12. This category received only 4 responses even though 9 participants chose this category as the one they belong to. The results show that 50% of the participants role involve designing graphics and/or user interfaces for Quranic applications. However, none use or apply secure design concepts while designing Quranic applications for sensitive

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

scripts such as Quran with one response (25%) was "never thought about this issue." Also, 50% didn't ever encounter any case in which the security of a digital Quran application was compromised due to unsecure design and 50% response was "never thought about this issue." As a result of the above responses all participants agree on the following points:

- They will consider and apply design security standards, techniques and measures to prevent any type of forgeries/tampering in the design of the Quran applications.
- It is necessary for a Quran body authentication body to monitor and endorse the digital copies/publications of the Quran worldwide.
- Prefer to validate Quran resources through a Quran authentication body.

Finally, when the participants were asked about providing comments/suggestions/vision as graphics designers only one response was received which suggested that social media can be linked to digital apps to encourage communications between people from all over the world can talk and discuss things about the Quran, also, there should be an option to contact scholars within such applications.

TABLE 12
THE "GRAPHIC DESIGNER" CATEGORY QUESTIONS

No.	Question	Response Count
1	Does your role involve designing Graphics or User Interface for Quranic applications?	4
2	Do you apply Secure Design concepts while designing Quranic application for sensitive Quranic Scripts?	4
3	Which Secure Design Concept do you use? (If Q.2 is Yes)*	3
4	Did you ever encountered any case in which the security of a Digital Quran application was compromised due to unsecure design?*	4
5	Will you consider and apply Design Security Standards, Techniques and Measures to prevent any type of forgeries/tampering in the design of Quran Applications?	3
6	Do you think it is necessary for a Quran Authentication Body to monitor and endorse the digital copies/publications of the Qur'an Worldwide?	4
7	Do you prefer validating your Quranic resources through a Quran Authentication Body?	4
8	What is your vision/comments/recommendations of how new technologies would be used to develop and improve Quranic/Related Smartphone Apps, Web Apps or user interaction further?	1

*Question has an "Other" option or "comment box"

D. Organization Manager/Director/Decision Makers/Owners/CEOs

This list of questions for the Organization Manager/Director/Decision Makers/Owners/CEOs category is provided in Table 13. The results show that approximately 60% of the participants are currently managers, directors, CEOs or decision makers of companies/organizations. Table 14 shows the current managerial role of the participants. The highest response "Mangers" option received 32.1% followed by "Other" option 29.5%. The "other" option included the following responses: deans of colleges, deputy deans, head of department, supervisors, professors, project managers, IT technologist, lecturers, and salespersons. 64.9% of the respondents are involved in Managing, Directing and Decision Making in Application Development. 75% of the companies work with smart devices and digital application development and only 15.3% of the companies develop and design digital Quran applications.

The results show that 65.2% of the digital Quran content is obtained from online free resources, 13% from Islamic bodies, 7.2% from Islamic institutes and 14.5% of the participants are not sure. In addition, 20.6% mentioned that they carry proof of authenticity for the Quran copies they develop without providing

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

any details, however, 79.4% do not have digital proof or never thought about Quran authentication. In response to the question "Does your company follow any particular security standards while developing these applications?" 34.3% responded "Yes" and 65.7% responded with "No" or "Never thought about using such standard." Table 15 provides the security standards followed by these companies. The "Other" option included ISO9001 + CMMI for DEV. L2, and SSL.

TABLE 13
THE "ORGANIZATION MANAGER/DIRECTOR/DECISION MAKERS/OWNERS/CEOS" CATEGORY QUESTIONS

No.	Question	Response Count
1	Are you the current Manager/Director/CEO/Decision Maker/Owner of your company/Organization?	76
2	What is your current Role?*	61
3	Are you involved in Managing, Directing and Decision Making in Application Development?	74
4	Does your company deal with Smart Devices/Digital Application Development?	72
5	Does your company develop and design Digital Quran Applications?	72
6	From where do you obtain the Digital Quran content?	69
7	What Proof-of-Authenticity do you have that the Digital Quran Copy is/was authentic?	68
8	Does your company follow any security standards while developing these applications?	70
9	Which Security Standards does your company follows? (If the previous answer is YES)*	56
10	Does your company follow Quality Assurance standards while developing such applications?	67
11	Which Quality Standards does your company follow? (If the previous answer is YES)*	50
12	Will you consider adopting Security Standards and Quality Assurance Standards while developing future digital applications?	66
13	Do you think that it is necessary for a Quran Authentication Body to monitor and endorse the digital copies/publications of the Qur'an Worldwide?	69
14	Will you prefer/recommend that your company will seek validating of Quranic resources through a Quran Authentication Body to obtain a Digital Quran Certificate?	70
15	What is your vision/comments/recommendations of how new technologies would be used to develop and improve Quranic/Related Smartphone Apps, Web Apps or user interaction further?	29

*Question has an "Other" option or "comment box"

TABLE 14
THE CURRENT MANAGERIAL ROLE OF THE PARTICIPANTS

Answer Options	Response Percent	Response Count
Manager	32.1%	25
Director	21.8%	17
CEO	9.0%	7
Decision Maker	6.4%	5
Owner	1.3%	1
Other please state)	29.5%	23

TABLE 15
THE SECURITY STANDARDS FOLLOWED BY THE COMPANIES OF THE PARTICIPANTS

Answer Options	Response Percent	Response Count
ISO 9001	17.9%	10
ISO 27001	3.6%	2
PCI-DSS	0.0%	0
FFIEC	0.0%	0
I don't know	71.4%	40
Other (please specify)	7.1%	4

In response to the question "Does your company follow Quality Assurance standards while developing such applications?" 41.8% of the participants responded "Yes", 25.4% responded "No" and 32.8% chose "I do not know". Table 16 provides the Quality Assurance standards followed by these companies. The "Other" option includes: BRC & ISO 9001, company standards, and WISE (in house developed method).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 16
THE QUALITY ASSURANCE STANDARDS FOLLOWED BY THE COMPANIES OF THE PARTICIPANTS

Answer Options	Response Percent	Response Count
ISO 9004	14.0%	7
ISO 19011	8.0%	4
AS9100	0.0%	0
MBNQA	0.0%	0
Lean/Six Sigma/TQM	0.0%	0
McCall Quality Factors	0.0%	0
I don't know	68.0%	34
Other (please specify)	10.0%	5

"Will you consider adopting Security Standards and Quality Assurance Standards while developing future digital applications" 79.4% say "Yes", 2.9% say "No" and 17.6% say "I don't know". A response with "I don't know" implies that such participants are not technically oriented or they do make decisions in the company.

"Do you think that it is necessary for a Quran Authentication Body to monitor and endorse the digital copies/publications of the Qur'an Worldwide?" 88.5 % responded Yes with high preference, 10.1 say "Not sure" and 1.4% say "No". In response to "Will you prefer/recommend that your company seek validating of Quranic resources through a Quran Authentication Body to obtain a Digital Quran Certificate?" 84.3% responded with "Yes", 12.9% "Not sure" and 2.9% "No". From the results, it is very clear that many of the participants in this category are not familiar with security and Quality Assurance standards.

Finally, the participants provided their vision/comments/recommendations of how new technologies would be used to develop and improve Quranic/Related Smartphone Apps, Web Apps or user interaction. The following list are the comments recorded by the participants,

- Multilanguage text search and translations.
- Validation of Quran from original source.
- Interactive voice commands
- Need to validate all Quranic information received from the Internet.
- Use data and text mining techniques
- Use NLP techniques as a supporting method in understanding the Holy Quran
- Provide clear, concise and uniform information on Islamic and Quranic issues.
- Authenticity is a must for Quranic applications.
- Recommend a single Islamic institution to develop and implement steps of policies, strategies, standards, procedures, technologies, tools, applications and implementations on international level.
- Support the development of an Islamic accreditation center which will have the authority to issue accreditation certificates for any Quranic/Islamic apps.
- There is an urgent need for authentic Quranic apps for new Muslims especially in Europe.
- Apply cloud computing and semantic web technologies as they are fertile fields for developing Arabic language and Quran apps, through research based on ontology relations and semantic fields in Arabic.
- Develop a matrix of basic standards for quality control and accreditation methodology for the Holy Quran Information Systems (HQIS) based on the "Quality Benchmarks" and "Accreditation" benchmarks. Develop a set of Key Performance Indicators (KPIs), which include a set of key factors and sub-functions. One of the systematic means of testing is to evaluate the basic stages of the Quran software in all its details and components, in order to improve the software and integrate IT in the service of Quranic sciences so that these techniques are integrated into the (HQIS). So, that to qualify

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the product to obtain a Quality Mark and to grant the manufacturer an Accreditation Certification, by a recognized donor, so that this matrix is the basis for anyone who wishes to produce software tools in the science of the Quran and is compatible with science and Sharia, whatever the form, type or structure of the program is.

- To ensure systematic dissemination of contemporary techniques used in the sciences of the Holy Quran in all its branches and elements, it is necessary to establish a comprehensive system of quality unified and acceptable from all specialized centers to evaluate and adopt the quality of the software tools related to the Holy Quran. Thus, combining the efforts between the researchers/workers in the field of IT, specialists in the forensic Quranic studies, experts in the field of international standards and software standards in order to develop a matrix of basic criteria for quality control and accreditation methodology for the software tools related to Quranic science. KPIs are used as one of the systematic means to test, measure and evaluate all the basic phases of the Holy Quran software in all its details and components. Therefore, the performance indicators are a set of key factors and sub-functions, which ensure quality controls and accreditation methodology for the Quran science software tools to improve them and integrate IT in the service of Quranic sciences, so that these techniques are integrated and linked to a vital link. To evaluate the accuracy, excellence and efficiency of these software tools, the quality controls and the accreditation methodology for the companies producing the Quran-related software tools will aid in achieving the required level of quality and meet the established standards regarding the organizational structure, procedures, processes and resources required to implement a comprehensive quality management system.

E. Islamic Institutes/Islamic Research Centers/Religious Body/IT or Other Research Centers

The list of questions for the "Islamic Institutes/Islamic Research Centers/Religious Body/IT or Other Research Centers" category is provided in Table 17.

TABLE 17
THE "ISLAMIC INSTITUTES/ISLAMIC RESEARCH CENTERS/RELIGIOUS BODY…" CATEGORY QUESTIONS

No.	Question	Response Count
1	Are you a representative of your institute/organization/center?	20
2	Does your curriculum cover Quranic studies at a Department or College level?*	17
3	Does your institute have Scientific Research related to Quran?	20
4	Does your institute have Research Centers dedicated for Quranic studies and research?	20
5	Does your institute address Digital Authentication and Certification issues in Quranic research and studies?	20
6	Do you think that the Digital Quran copies/publications available online or on different devices are all100% authentic?	19
7	Based on the Quranic research and academic activities conducted in your institute, do you think that itis necessary for a Quran Authentication Body to monitor and endorse the digital copies/publications of the Qur'an Worldwide?	20
8	Will you prefer that your institute will become a part of such initiatives to secure and validate the digital Quranic resources by forming a Digital Quran Authentication Body and issue a Digital Quran Certificate that confirms the authenticity of the Quran contents?	20
9	What is your vision/comments/recommendations of how new technologies would be used to develop Quranic/Related Smartphone Apps, Web Apps or user interaction further?	10

*Question has an "Other" option or "comment box"

This category received 20 responses, 85% of the participants are representative of their institutes/organizations however, 15% are representatives of non-Islamic institutes/organizations. 85% of these institutes/organizations provide Quranic studies curriculum at a college or department level and 15% at a high school or grade school level. Table 18 provides the responses to 5 of the questions shown in Table 17. The results show that 65% of the research in these institutes is dedicated to Quran with 55% of these

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

institutes having special research centers for Quranic studies. However, even though most of these research institutes carry research on such sensitive script the issue of authentication is not given high importance, since only 25% of these institutes address the issue of digital authentication and certification. The 15.8% of the respondents believe that the digital copies of the Quran available online or on different devices are 100% authentic. Therefore, 80% of the participants prefer that their institute will become a part of such initiatives to secure and validate the digital Quranic resources by forming a Digital Quran Authentication Body and issue a Digital Quran Certificate that confirms the authenticity of the Quran contents.

TABLE 18
THE RESPONSE TO SOME OF THE QUESTIONS IN THIS CATEGORY

Question	Yes	No	Not sure
Does your institute have Scientific Research related to Quran?	65%	35%	0%
Does your institute have Research Centers dedicated for Quranic studies and research?	55%	45%	0%
Does your institute address Digital Authentication and Certification issues in Quranic research and studies?	25%	55%	20%
Do you think that the Digital Quran copies/publications available online or on different devices are all 100% authentic?	15.8%	21.1%	63.%
Will you prefer that your institute will become a part of such initiatives to secure and validate the digital Quranic resources by forming a Digital Quran Authentication Body and issue a Digital Quran Certificate that confirms the authenticity of the Quran contents?	80%	15%	5%

Finally, from the comments of the participants it can be concluded that:

- Apps should provide certification and guidance to users.
- Apps should be designed with age groups in mind.
- Apps should be user-friendly
- Authentication is a must.
- Develop Authentication filters to check all information on web.
- ISO certification.
- Use such applications in teaching the Holy Quran.

F. Holy Quran Publishers

The list of questions for the Holy Quran Publishers is given in Table 19. It shows that the maximum number of responses recorded for the Holy Quran Publishers category is 4. The results show that all participants agreed on the following points:

- Their companies publish and distribute Quran related digital content or printed copies in addition to other non-Quranic publications.
- Publication of digital Quran in different media such as smart devices and online are effective and easy methods for users to read Quran.
- Current digital copies of the Quran are free from any intentional or intentional errors.

Other results from the responses received include:

- 66.7% think that the digital copies of the Quran are all 100% authentic and the participant's organizations can rely on its authenticity; however, 33.3% are not sure.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- In responding to the question "Will your publishing organization consider publishing digital copies of the Quran?" three responded to this question and each one chose a different response, these are: "Yes", "No", and "I do not know."
- Most of the participants (3 out of 4) are unaware if any authentication body is equipped and trained enough to conduct such authentication and validation of digital Quran/resources/publications on the other hand only 1 respondent agreed on this.
- 75% of the 4 participants agree that it is a very high preference/requirement for digital copies of the Quran to be monitored and endorsed by a Quran authentication body and 25% thinks otherwise.
- 66.7% (3 out of 4 participants) would prefer/recommend that their publishing company will consider applying for validating their Quranic resources through a Quran Authentication Body to obtain a Digital Quran Certificate that confirms the authenticity of the digital Quran contents.

TABLE 19
THE "HOLY QURAN PUBLISHERS" CATEGORY QUESTIONS.

No.	Question	Response Count
1	Does your company deal with Publishing and Distributing the glorious Quran? (Digital or Printed)	3
2	Does your company ONLY deals with Publishing and Distributing the printed and non-digital copies/publications of Quran?	3
3	Is there any Authentication Body that deals in authenticating and validating the Quran before final publishing and further distributing?*	3
4	Does your company consider that Digital Quran available in Smart Devices or Online provides another effective and easy method for users to read Quran?	3
5	Do you think that the current digital copies of the Quran is free from any intentional or unintentional errors?	3
6	Do you think that the digital copies of the Quran are all 100% authentic and your publishing organization can rely on its authenticity?	3
7	Will your publishing organization consider publishing digital copies of the Quran?	3
8	Do you think that any Authentication Body is equipped and trained enough to conduct such authentication and validation of digital Quran resources/publications?	4
9	Do you think that it is necessary for a Quran Authentication Body to monitor and endorse the digital copies of the Qur'an Worldwide?	4
10	Will you prefer/recommend that your publishing company will consider applying for validating their Quranic resources through a Quran Authentication Body to obtain a Digital Quran Certificate that confirms the authenticity of the digital Quran contents?	3
11	What is your vision/comments/recommendations of how new technologies could be used to develop and improve Quranic/Related Smartphone Apps, Web Apps or user interaction further?	1

*Question has an "Other" option or "comment box"

Finally, when asked for their comments, recommendations of how new technologies could be used to develop and improve Quranic related smartphone apps, or web apps only one participant suggested to apply Artificial Intelligence for understanding the Quran.

IV. Recommendations

Based on the comments, suggestions and recommendations provided by the participants in different categories surveyed, the main recommendations can be summarized as follows:

- Establish an international Islamic accreditation center to be the only source for publishing the printed and digital copies of the Quran. Its role will also include training, monitoring, endorsing, and scrutinization.
- Support the development of an Islamic accreditation center which will have the authority to issue accreditation certificates for any Quranic/Islamic center so that they can act as certification bodies under the umbrella of the main Islamic accreditation center.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- The International Islamic accreditation center must set a framework with clear milestones for developing standards the issuance of certifications for centers as well as for the design of web and mobile Quran apps.
- This establishment/organization is also responsible to provide certificates of authorization to establish centers in different countries that follow strict rules in publishing Quran printed copies and issue certifications for apps.
- Develop a matrix of basic standards for quality control and accreditation methodology for the Holy Quran Information Systems (HQIS) based on the "Quality Benchmarks" and "Accreditation" benchmarks.
- Develop a set of Key Performance Indicators (KPIs), which include a set of key factors and sub-functions. One of the systematic means of testing is to evaluate the basic stages of the Quran software in all its details and components, in order to improve the software and integrate IT in the service of Quranic sciences so that these techniques are integrated into the (HQIS). So, that to qualify the product to obtain a Quality Mark and to grant the manufacturer an Accreditation Certification, by a recognized donor, so that this matrix is the basis for anyone who wishes to produce software tools in the science of the Quran and is compatible with science and Sharia, whatever the form, type or structure of the program is.
- Apply cloud computing and semantic web technologies as they are fertile fields for developing Arabic language and Quran apps, through research based on ontology relations and semantic fields in Arabic.
- Develop a secure cloud-based repository for Quran related resources and make it available as a web-service for all mobile and web apps. Any Quran related resources must be first authentic before being used in designing apps.
- Provide in app capability for real time authentication.
- Develop an agent program to keep monitoring the Quran text for correctness. Such text may be any uploads related to Quran with the help of popular social media and other websites that may contain such contents.

In regard to features the participants would like to see in mobile and web apps, the following list provides some of the ideas presented by them:

- Social media can be linked to digital apps to encourage communications between people from all over the world can talk and discuss things about the Quran.
- Option to contact scholars within the app.
- Develop interactive apps.
- Provide in-app technical support for monitoring and addressing concerns from users.
- Develop an intelligent Quran content checker.
- Anti-tampering feature should be included in such applications to alert the users.
- Offline accessibility.
- Provide free of charge apps and ad free apps
- Encapsulates different features in one app such as qibla direction, prayer time … etc.,
- User friendly voice-driven apps.
- Develop apps targeting people with special needs.

V. Conclusion

This study presented the results of a questionnaire survey distributed to 500 participants to investigate the aspects of security and authentication, certification and standardization of Quran related applications. The survey provided the feedback/perspective of six different categories of participants and this is the first of its

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

kind study/survey carried out in Quran computing to confront the issues of authentication, certification and standardization. The participants pointed out that there are few incidents that confirm the existence of errors in few digital versions of the Holy Qur'an that are deployed via the Internet and/or mobile apps and such intentional or unintentional forgery is not acceptable even if it is a small error such as a removal or addition of one letter or diacritic, thus this will deem the app or digital content distorted and invalid. Therefore, it can be concluded that the main concern is not the development of Quran apps with attractive features or cosmetic user interface. Certainly, the issue is the available and standardization of properly/accurately digitized Quran related content/resources that are well authenticated (by scholars) and widely distributed (by recognized bodies). Then the risks in terms of security aspect would be very much reduced and controlled. If this stage is reached then, it will be very efficient and easy to develop sophisticated and advanced apps that can be easily monitored and certified by recognized bodies. The latter is the quality standard that we need to primarily develop for Quran related apps, rather than any other standard that would check/improve the technical aspect of an app. Finally, security and authentication of digital Quran increases the confidence of the user in using Quran apps.

ACKNOWLEDGEMENT

The authors would like to thank and acknowledge the King Abdulaziz City for Science and Technology (KACST) for their financial support for this research work during the academic year 2015/2016 under research grant number "440-34 - ت ا". Also, the authors would like to thank the IT Research Center for the Holy Quran and Its Sciences (NOOR) at Taibah University for providing the facilities to conduct research.

REFERENCES

[1] Khan, Muhammad Khurram, and Yasser M. Alginahi. "The holy Quran digitization: Challenges and concerns." Life Science Journal 10.2 (2013): 156-164.
[2] Tayan, O. and Alginahi, Y., "Information and Communication Technologies in the Teaching and Spreading of the Holy Quran", Symposium on the Glorious Quran and Contemporary Technologies, Information Technology, October 13-15, Madinah, Saudi Arabia, 2009.
[3] Tayan, Omar, and Yasser M. Alginahi. "A review of recent advances on multimedia watermarking security and design implications for digital Quran computing." In Biometrics and Security Technologies (ISBAST), 2014 International Symposium on, pp. 304-309. IEEE, 2014.
[4] Tayan, Omar. "Concepts and Tools for Protecting Sensitive Data in the IT Industry: A Review of Trends, Challenges and Mechanisms for Data-Protection." International Journal of Advanced Computer Science and Applications 8, no. 2 (2017).
[5] Hakak, S., Kamsin, A., Tayan, O., Idris, M. Y. I., Gani, A., & Zerdoumi, S. (2017). Preserving Content Integrity of Digital Holy Quran: Survey and Open Challenges. 7, IEEE Access. DOI: 10.1109/ACCESS.2017.2682109.
[6] Zakariah, Mohammed, Muhammad Khurram Khan, Omar Tayan, and Khaled Salah. "Digital Quran Computing: Review, Classification, and Trend Analysis." Arabian Journal for Science and Engineering (2017): 1-26.
[7] Sabbah, Thabit, and Ali Selamat. "A framework for Quranic verses authenticity detection in online forum." In Advances in Information Technology for the Holy Quran and Its Sciences (32519), 2013 Taibah University International Conference on, pp. 6-11. IEEE, 2013.
[8] Alsmadi, Izzat, and Mohammad Zarour. "Online integrity and authentication checking for Quran electronic versions." Applied Computing and Informatics13, no. 1 (2017): 38-46.
[9] Mostafa, Mostafa GM, and Ibrahim M. Ibrahim. "Securing the digital script of the Holy Quran on the Internet." In Advances in Information Technology for the Holy Quran and Its Sciences (32519), 2013 Taibah University International Conference on, pp. 57-60. IEEE, 2013.
[10] Hilmi, Mohd Faiz, Mohammed Fadzil Haron, Omar Majid, and Yanti Mustapha. "Authentication of electronic version of the Holy Quran: an information security perspective." In Advances in Information Technology for the Holy Quran and Its Sciences (32519), 2013 Taibah University International Conference on, pp. 61-65. IEEE, 2013.
[11] AlAhmad, Mohammad A., Imad Alshaikhli, and Bashayer Jumaah. "Protection of the Digital Holy Quran hash digest by using cryptography algorithms." In Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on, pp. 244-249. IEEE, 2013.

[12] Mohammed S. Khalil, Fajri Kurniawan, Muhammad Khurram Khan, and Yasser M. Alginahi, "Two-Layer Fragile Watermarking Method Secured with Chaotic Map for Authentication of Digital Holy Quran," The Scientific World Journal, vol. 2014, Article ID 803983, 29 pages, 2014. doi:10.1155/2014/803983

[13] Kurniawan, Fajri, Mohammed S. Khalil, Muhammad Khurram Khan, and Yasser M. Alginahi. "DWT+ LSB-based fragile watermarking method for digital Quran images." In Biometrics and Security Technologies (ISBAST), 2014 International Symposium on, pp. 290-297. IEEE, 2014.

[14] Khan, Muhammad Khurram, Zeeshan Siddiqui, and Omar Tayan. "A secure framework for digital Quran certification." In Consumer Electronics (ICCE), 2017 IEEE International Conference on, pp. 59-60. IEEE, 2017.

Proposed Architecture for Improving Security and Consistency of Data Transactions in Cloud Database using Tree-Based Consistency Approach

J.Antony John Prabu,
Research Scholar and Assistant Professor,
Department of Computer Science
St. Joseph's College, Trichy, Tamilnadu, India- 620 002

Dr.S Britto Ramesh Kumar
Assistant Professor,
Department of Computer Science
St. Joseph's College, Trichy, Tamilnadu, India- 620 002

***Abstract*— Cloud computing is one of the important resources in IT industries to provide different services for cloud vendors and clients. Security is the main issue in cloud because data are stored and maintained by third party environment. Cloud computing has lot of issues to maintain transactional data in cloud databases. It needs to maintain ACID guaranties to execute the transactional data. In this paper, Proposed Architecture, Cloud controller, Depth 1 Fixed Tree Consistency (D1FTC) method, Cloud Data Locker and data storage in cloud have been discussed. It also portraits the needs of ACID guaranties, major security levels and ensure consistency of data transactions. Finally this paper affords thorough study on proposed architecture for improving security and consistency of data transactions in cloud database.**

***Keywords*--- Cloud DTM, 2-Phase Commit Protocol, Cloud Security, Consistency in cloud DBS, Cloud Controller, D1FTC and Cloud Data Locker.**

I. INTRODUCTION

Cloud computing provides everything as a service and it has different deployment models. It gives reliable service for analytical data but not for transactional data management [32]. Most of the cloud services are deployed in the hybrid cloud environment. The cloud vendors are making a service agreement with cloud providers and launch the reliable services for their clients. Instead of owning the software and hardware vendors move to cloud environment. So the cloud vendors need not worry about the software and hardware maintenances. Cloud has some technical issues to handle the data in distributed environment [10]. The cloud researchers are designing and developing new techniques for scalable transactions to avoid workload consistent[47,49], concurrency controls to improve serialization access [4, 8, 22], consistency approaches for efficient transaction processing system[24,25,26], scalable data storage model for better data handling[7,12,14], data security models[18,19,21] for efficient data handling and security algorithms for multilevel security. Therefore providing security is one of the common and major concerns in cloud, so multilevel security is needed to product each service without loss of data [17, 33, 48]. The next major issue is maintaining strong consistency state in the database level especially for transactional data management services [46]. Hence most of the research works are concentrating to develop the reliable methodology and architectures to strengthen the security for cloud services and strengthen consistency state in cloud distributed database system [15, 16, 30].

II. REVIEW OF LITERATURE

R.Anandhi, et all proposed a model to improve the consistency state of the data transaction and also analyze the performance factors of different scalability options in cloud databases. This paper insists that reliability of cloud transaction applications depend on the consistency and scalability levels [13]. Chang Yao, et all proposed a concurrency control protocol named DGCC ((Dependency Graph based Concurrency Control). This protocol is used to achieve better execution of transaction and scalability among database systems. This DGCC based OLTP system also integrated with efficient recovery mechanisms [1]. Pornpan Ampaporn, et all explore a performance analysis against two different data consistency models by leading cloud providers. The results show that writes performance was 3 times worse than reads and it also has grater variance of consistency rate. Hence a better consistency approach is needed for data transactions in cloud environment [3]. Aleksey Burdakov, et all proposed a consistency model for NoSQL databases for data transactions. It explores the characteristics of data consistency and analysis performance of different consistency models [6]. Álvaro García-Recuero, et all proposed a consistency model to efficiently replicate the data among long geographic distance in cloud environment. This approach secures the overloading of both network and system side. The architecture builds with three dimensional vector field models to handle different applications in cloud [5]. Jens Kohler, et all proposed a architecture called data cache architecture with implementation of both parallel and lazy fetch strategies. This work explores the performance analysis between the two strategies and discusses to overcome the issues with SeDiCo framework [7]. Thuy D. Nguyen, et all proposed a prototype called MLS column-store following kernelized design pattern. This approach used in cloud-scale data storage system. It explores the guaranties of efficient cloud-scale data storage in distributed system [17]. Sebastijan Stoja, et all proposed a architecture for realtime database in cloud data transactions. This paper explores the other important topics and analyzes the merits and limitations of it [38]. Marco Serafini, et all proposed

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

a dynamic data placement system for partitioned database management system. This system provides ACID guaranties for data transactions. The analyzed results are shown the server capacity and it used to improve the placement quality [38].

III. NEED OF ACID

Ensure guarantee to maintain ACID properties in data transaction is global rule in IT sector. The service providers design efficient architectures to satisfy this issue [27]. Especially in cloud, data are stored in third part environment maintained by cloud vendors and cloud providers [34]. Hence maintaining ACID in cloud is not trivial [28], it is a complex work to implement efficient data transaction services in cloud. Many cloud providers offer their service for data transactions but they have their own merits and limitations [34,35].

The main idea of transaction is sequences of data read and write.

TABLE I. ACID PROPERTIES

A	A transaction executed completely or It goes to the initial state. [All or Nothing]
C	Maintain Consistent state in database
I	The particular transaction will not affect other transactions
D	Ensure a transaction committed successfully

When a transaction committed successfully in cloud environment, it should satisfy the ACID guaranties without loss of any details. The performance analyses depend on the ACID maintenances [36].

IV. EXISTING CONSISTENCY APPROACHES

The frequently used recent consistency approaches in the distributed cloud environment are Classic approach, Quorum approach and Tree-Based approach. The through study of these three consistency approaches are discussed with its merits and limitations.

a) Classic Approach:

The classic approach maintains consistency state through synchronous replication in distributed environment. In this approach all nodes or database servers have participated in the writing operation [2]. Hence, it has low consistency rate and big execution time for each transaction in cloud environment [9, 23].

b) Quorum Approach:

The Quorum approach is stronger than the classic approach that is frequently used in cloud to replicate the cloud databases. The so-called quorum method of voting is used for replication among the cluster of servers in cloud environment. In this approach, the majority of votes from all participated nodes or servers are confirmed for further execution. It shows more performance and gives high consistency assurance than the classic approach, but it slow down during the database execution because of frequent voting system [9, 23].

c) Tree-Based Approach:

The Tree-Based consistency approach is the leading approach to maintain data consistency in the distributed data bases. It is formulated based on complex tree structure. Hence it provides varies degree of consistency rate depends on the replica server placed in the level of trees[2]. This approach introduces components in the cloud environment for the better execution of transactions [9]. The highest level of replica nodes in the tree provides high consistency assurance but it gives low consistency assurance moves to lowest level of replica nodes on the tree [9, 23].

V. SECURITY AT DIFFERENT LEVELS

Multi level security is essential for distributed database system. Every level of protection is important to avoid data loss for efficient data service. More concerns are needed to provide data service in cloud because data are maintained by third party environment [37, 45]. So proper security models are may strengthen the cloud services [20,31].

Cloud services need security at following levels:

TABLE II. V. SECURITY AT DIFFERENT LEVELS

Level of Security	Description
Server access security	It ensures the access control (Authentication, Authorization, and Auditing) to services in the cloud environment.
Internet access security	Connectivity and Open access manage in the public cloud. Infrastructure Security at the Network Level. Ensuring availability of the Internet facing resources of the public cloud used by the organization.
Database access security	Ensuring Access control for database and Key management for encrypting.
Data privacy security	Ensuring data confidentiality and integrity of the organizations data in transit to and from the public cloud provider.
Program access Security	Ensuring access control security for the programs of the client's applications in the public cloud.

VI. DATA TRANSACTION IN CLOUD

a)ElasTraS: An Elastic Transactional Data Store in the Cloud:

ElasTraS is designed for scalable and elastic data transaction in cloud databases[39]. It add components to achieve the elasticity in data storage during data transaction.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

It uses two level hierarchy to maintain transaction guarantee and also make elastic scalability while increasing workload. ElasTraS has overcome the limitations of DDBMS with its database techniques for isolation and concurrency control.

b) G-Store: A scalable Data Store for Transactional Multi key Access in the Cloud:

G-Store is a scalable data store with multi key access in the cloud environment [40]. It is designed to achieve scalability, availability and fault-tolerance. The key group abstraction procedure is allowed to select any set of key group in the data store and provide scalable transactions. The atomicity and consistency guarantees are maintained by the single key to group of keys .

c) Scalable Transactions for Web Applications in the Cloud:

This approach supports scalable transactions in cloud environment [29] . It has transaction manager and many number of local transaction managers to handle the transactions. It maintains the ACID properties even in the server failures. The local transaction managers replicate the data and periodically checkpoint data snapshots to cloud data storage service [30].

d) EcStore: Towards Elastic Transactional Cloud Storage with Range Query support:

The EcStore deployed among the cloud cluster to provide high elasticity with efficient range query to support cloud data transactions [41]. It achieves the features like load balancing, data partitioning, data replication and efficient range query for each transactional access. The distributed storage layer, replication layer and transaction manager layers are supported to handle data in cloud storage system.

e) Dynamo: Amazon's Highly Available Key-Value Store:

Dynamo is built for amazon to achieve high availability and scalability among the cloud clusters [42]. It takes a step to satisfy high availability, consistency, performance and cost-effectiveness. In this approach data is partitioned and replicated with consistent hashing and consistency is maintained by object versioning. The quorumlike technique is used to maintain consistency state among replicas during transaction updates with decentralized replica synchronization protocol.

f) Megastore: Providing Scalable, High Available Storage for Interactive Services:

Megastore is a highly available storage system for interactive services in cloud [43]. Most of the NoSQL storage system like Google's Bigtable and HBase are fully support scalable but they have loose consistency model to maintain consistency state. It satisfy ACID properties over remote replicas with low latency for interactive applications.

g) Sinfonia: a new paradigm for building scalable distributed systems:

Sinfonia provides efficient and consistent access of data for mini data transactions to abstract the problem faced from concurrency and transaction failures [44]. It avoids the message passing protocols to minimize the complexity for the development process. The Sinfonia developers manipulate data centre infrastructure system like file system, lock manager and communication services.

VII. PROPOSED ARCHITECTURE

The proposed architecture mentioned in Fig improves security and consistency for data transactions in cloud environment. The components placed in this architecture are discovered after analysis of related works done by the field.

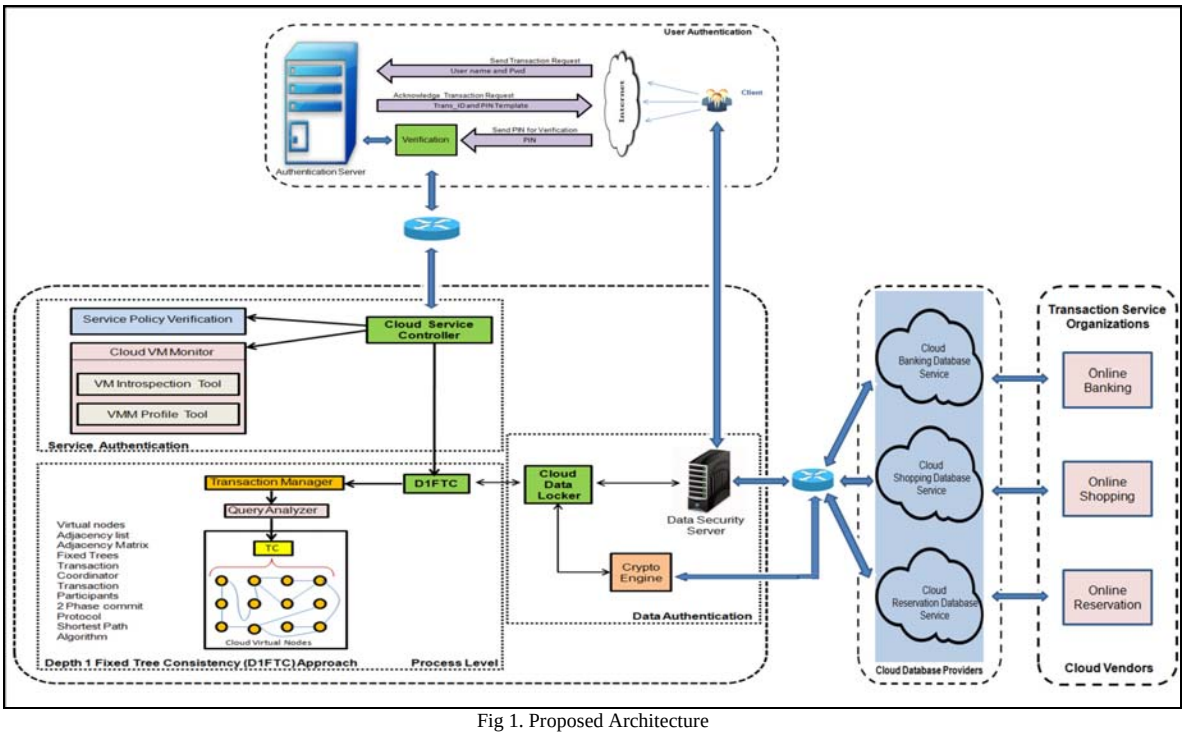


Fig 1. Proposed Architecture

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The further execution details are explored thoroughly in this section. main components are mentioned shout.

a)User Authentication

b) Cloud Service Controller

c) Depth 1 Fixed Tree Consistency (D1FTC) Approach

d) Cloud Data Locker

a)User Authentication:

The client sends the transaction request to authentication with username and password through internet. The authentication server acknowledges the request and sends the transaction ID with PIN template to the client. The client sends the PIN to authentication server for verification. The authentication server verifies whether it is a correct PIN for the specified account or not. After the proper verification process the request may enter to the service level.

b) Cloud Service Controller

The frame work for the cloud service controller makes clear the functionalities of it to enhance the service authentication. After the user authentication the client's requests enter to the cloud service controller. It has two main components named service policy verification and cloud virtual machine monitor.

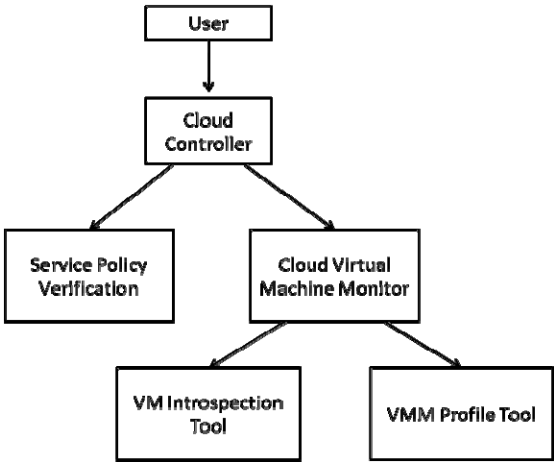


Fig 2. Cloud Service Controller

b.1) Service Policy Verification:

The service policy verification verifies the SLA between cloud providers and cloud vendors. It ensures the Measurement, Condition Evaluation and Management services are not violated. After the service policy verification the request enters in to the virtual machine monitor.

b.2) Cloud Virtual Machine Monitor:

It has two different tools named VM Introspection and VMM Profile tools. Mainly these tools are occupied to verify the status of the each virtual machine in the data transaction management.

b.2.1) VM Introspection Tool:

Number of virtual machines are created and involved in the data transaction execution. The virtual machine introspection tool especially used to evaluate the virtual machines. It frequently verifies the virtual machines and ensures the virtual machine functionalities are good. This approach is very much useful to maintain the virtual machines for providing reliable service for data transaction clients.

b.2.2) VMM Profile Tool:

Virtual Machine Manager (VMM) contains some of the specifications about the virtual machines. It has virtual machine profiles used to simplify the work to create templates. Templates are used to create virtual machines quickly with proper hardware and operating system settings. The VMM profile has different types of profiles like hardware profile, operating system profile and virtual hard disk profile. These profiles are used to create virtual machine from the created templates.

c) Depth 1 Fixed Tree Consistency (D1FTC) Approach:

The proposed D1FTC method is especially comfort with data transactions in cloud. It efficiently supports 2 phase commit protocol to execute each transaction without affect ACID properties even in the critical situations. The preliminary setup and methodology of the proposed D1FTC method is elaborately explained as follows.

c.1) Functional Procedures:

Preliminary setup:

Step 1: Generate an undirected graph for on hand virtual machines (nodes)

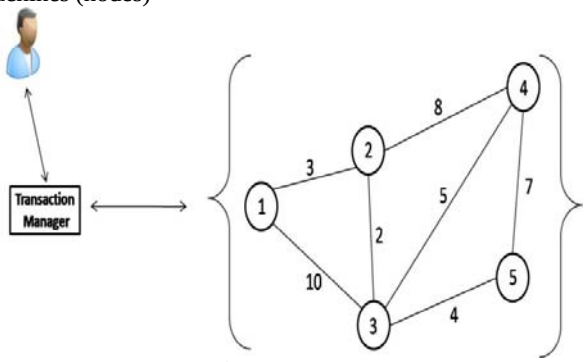


Fig 3. undirected graph for cloud virtual machines with distances

Step 2: Create possible Depth 1 Fixed Trees (D1FTC)

Create adjacency list for the undirected graph and Refer it to find possible depth 1 fixed trees. The number of nodes in undirected graphs is considered as cloud virtual machines.

Adjacency list for the graph is as follows:

Node	Possible Linked Nodes
1	2 3
2	1 3 4
3	1 2 4 5
4	2 3 5
5	3 4

Fig. 4. Adjacency List

The adjacency list specifies the relationship between each node to other nodes in the graph. Hence it very much useful to create possible number of depth 1 fixed trees for the graph as follows:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

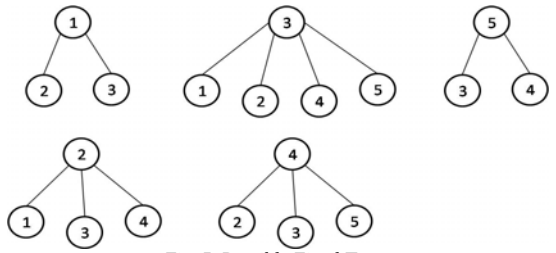


Fig. 5. Possible Fixed Trees

Step 3: Find the distance between each nodes in D1FTC

Create adjacency matrix for the undirected graph and refer it to find the distance between nodes in the tree and count the nodes in each trees. The adjacency matrix for the above mentioned undirected graph is as follows:

	1	2	3	4	5	No. Of Nodes	Distance between Nodes
1	α	3	10	α	α	3	13
2	3	α	2	8	α	4	13
3	10	2	α	5	4	5	21
4	α	8	5	α	7	4	20
5	α	α	4	7	α	3	11

Fig. 6. Adjacency Matrix

Step 4: Find the shortest path from all nodes

The objective is to find the shortest path from each node to all other nodes in the graph. The Dijkstra's Algorithm is suitable way to find the shortest path from one node to other nodes in the graph.

Fig (a), desire to find shortest path from node 1. Edge values in the graph are weights and Node values in the tree are total weights. The transaction manager fixes one nearest node for starting position to find shortest path to connect all nodes. Fig (b) is the shortest path from node 1 and Fig (c) is created as fixed tree for shortest path. It works to update the data among all cloud virtual machines after successful commit of the cloud transaction. So the transaction manager replicated the data after successful commit with shortest path of fixed trees.

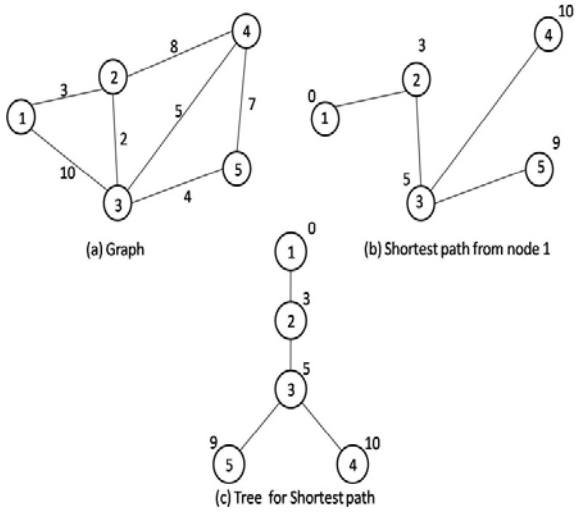


Fig. 7. Find shortest path

D1FTC Method:

The D1FTC method is ready for execution after the proper preliminary works. It handles the virtual machines efficiently to commit each transaction successfully.

The components of D1FTC are as follows:

1. User
2. Transaction Manager
3. Fixed Trees
4. Transaction Coordinator
5. Transaction Participants
6. Two-Phase commit protocol
7. Shortest path Trees

a) User:

The user can submit request to the transaction manager and get response without loss of details.

b) Transaction Manager:

The transaction manager manages all transaction requests from different users and it fix a node for read query and fix a tree for update operations. It analyzes the query and fix the suitable tree with transaction coordinator and participants.

c) Fixed Trees:

The structures of these possible fixed trees are reliable to implement two-phase commit protocol, because it has one to many relationships that is one transaction coordinator and many transaction participants for all update transactions. A participant need not affect the other participants and it communicates only with the coordinator. So the transaction manager can choose any one of the node for transaction coordinator and linked nodes under the transaction coordinator are chosen as transaction participants.

c.1) Adjacency Matrix:

The transaction manager fixes any one of the tree that depends on the required nodes needed to execute a transaction. So the adjacency matrix calculates the number of nodes and the distance between nodes that is to simplify the work of transaction manager.

d) Transaction Coordinator:

Transaction Coordinator is responsible for the given transaction and it maintains all the participants in the selected fixed tree to commit a transaction. The two-phase commit protocol implements in the transaction coordinator.

e) Transaction Participants:

The transaction is divided into small no of process and it sends to transaction participants in the fixed tree. All participants are under supervised by the transaction coordinator.

f) Two-Phase commit protocol:

Two-Phase commit protocol is one of the efficient ways to execute data transactions in the distributed system. It can help successfully and execute the transactions with ACID guaranties in cloud environment. It is also very reliable for the proposed D1FTC method.

g) Shortest path Trees:

In cloud, data replicates from large geographic distance for every transaction and data may be lost during the replication process. The work of the finding shortest path is to connect all

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

virtual machines. It is to avoid the inconsistency database and minimize the replication time in cloud environment.

c.2) Framework:

The frame work for D1FTC approach is elaborately shown bellow. The step by step process executions are clearly defined and it leads to execute the data transaction and replicate the data in cloud distributed database system.

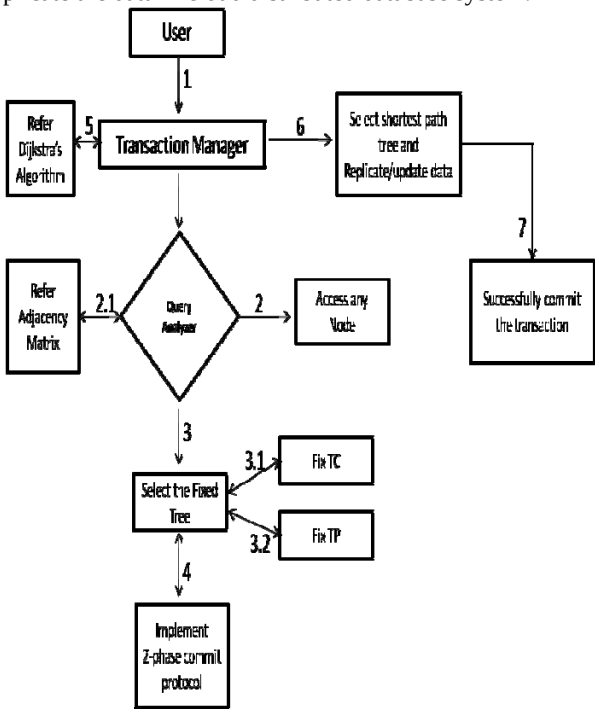


Fig. 8. Frame work for D1FTC

c.3) Pseudocode for D1FTC Approach :

Step 1 : User request sends to transaction manager

Step 2 : If (Read operation)

{

Allocate any node to get updated data.

}

Else

Step 2.1:Refer Adjacency Matrix

Step 3 : Select the Depth 1 Fixed Tree [According to the transaction query,]

Step 3.1:Fix the root node as transaction coordinator(TC)

Step 3.2:Fix the child nodes are transaction participants(TP)

Step 4 : Implement 2-phase commit protocol in the TC.

TC divides the transaction and send it to Transaction Participants (TP)

TC executes the transaction

Step 5 : Refer Dijkstra's Algorithm

Step 6 : Select shortest path tree and Replicate/update the data.

Step 7 : Successfully commit the transaction

d) Cloud Data Locker

In cloud, data are stored in the third party environment and it accessed by remote clients. The cloud providers offer database as a service with their own functionalities. The cloud vendors purchased the Infrastructure, platform, software and database and launch the services for their clients. Hence maintaining security in data storage level is not easy for cloud data transactions. The proposed cloud data locker model has three stage verification.

d.1) Frame work:

The first stage Cloud Data Locker (CDL) model verifies the user to send and get OTP through internet. The second stage it accesses Data Security Server and it verifies the data storage (providers) to send and get the Data Security Number (DSN). The third stage is enabled after the successful verification of both stage one and two. It accesses the Crypto Engine to handle the encrypted data stored in cloud data storage. The crypto engine decrypts the data for execution and encrypts the data after execution to store it in cloud data storage. So the provider does not know which data stored by the clients and the user, vendors are verified with OTP and DSN systems. Hence these three stage verification ensure the high data level security for cloud environment.

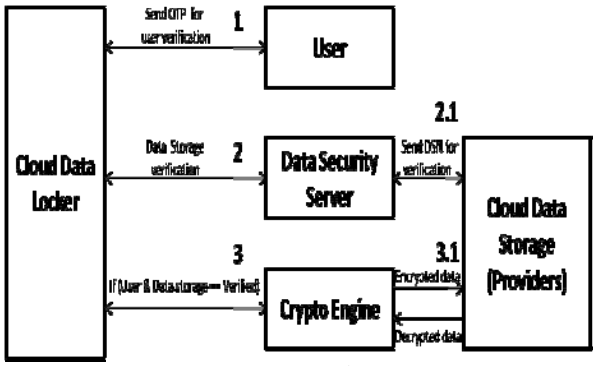


Fig. 9. Frame work for CDL

d.2) Pseudocode for cloud data locker:

Step 1: Send OTP to user

Step1.2: get OTP from user

Step1.3: Verify the user

Step2: Send Data Security Number (DSN) to Cloud Data storage provider

Step2.1: Data Security Server match the DSN in Data storage provider

Step2.2: Verify the Cloud Data storage provider

Step3: If (User & Data storage == Verified)

Step3.1: Access the Crypto Engine and decrypt the data

Step3.2: Send data to the transaction process

Step3.3: Get the committed data

Step3.4: Access the Crypto Engine and encrypt the data

Step3.5: Store the data successfully.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VIII. FEATURES OF PROPOSED ARCHITECTURE

a) Security for all levels

The proposed architecture provides multilevel security for every data transactions. User authentication server, cloud controller, cloud data locker are constructed to strengthen the security in user, service and data storage levels.

b) Ensure Service level agreement:

The SLA between cloud providers and vendors are verified continuously to offer reliable service for clients of the vendors. The service policy verification makes sure the SLA is up-to-date.

c) Efficient Consistency Method:

D1FTC is a proposed consistency approach designed for data transaction management in cloud. The adjacency list and adjacency matrix are referred to construct depth 1 fixed trees. It measured the number of nodes and distance between nodes in a fixed tree. After every commit, data are replicated through the shortest path tree referred with Dijkstra's Algorithm.

d) Consistent structure for 2-phase commit protocol:

The depth 1 fixed trees are structured specially to implement two-phase commit protocol. All fixed trees have one transaction coordinator and more than one immediate transaction participants. Hence the D1FTC structure is fully reliable for two-phase commit protocol.

e) Enhanced Data Security:

The proposed Cloud Data Locker model has three stage verification processes to enhance the data storage security in cloud. It verifies the user authentication with OTP system, verify transaction services with data security server and access the data through crypto engine. The data are stored in encrypted form in the cloud. The cloud providers and vendors cannot access or operate the client's data.

f) Guarantee Access Control:

The proposed architecture ensures the access control in all levels. It grants proper implementation of user authentication server, cloud service controller and cloud data locker to manage access control in user, service and data storage level.

g) Ensure ACID guaranties:

The proposed D1FTC approach maintains the *Consistency* state for data transactions in cloud. The two-phase commit protocol is implemented in the fixed trees of D1FTC approach to maintain the *Atomicity* and *Isolation* properties. After the data transaction execution data are updated / replicated through the shortest path tree to maintain the *Durability* of every data transactions. So the proposed architecture ensures the ACID guaranties for data transactions in cloud environment.

h) Efficient data replication:

After every execution of transaction the data should update/ replicate with all servers in distributed clod database. The proposed D1FTC refer the Dijkstra's Algorithm to find the shortest path tree from each node to all other nodes in cloud. The transaction can choose reliable shortest path tree to update / replicate the data efficiently with in the short span of time compare with other approaches.

i) Trusted Transactions:

The cloud service providers (vendors) and end users expect conviction for every transaction. The proposed D1FTC approach has transaction manager to track the transaction from beginning to go to commit state. The transaction coordinators in fixed trees are responsible for allotted transactions. If the system faces any failure, it goes to the initial state without loss of the data.

j) Easy implementation:

The implementation of proposed architecture is simple in the cloud environment. The virtual machines are fixed for transaction process with the reference of Adjacency list, adjacency matrix and Dijkstra's Algorithm in the proposed D1FTC approach. The proposed cloud service controller has the VM introspection and VMM profile tolls to monitor the state of virtual machines and maintains them. The proposed cloud data locker has simple three stage methodology to provide high level of data security in the cloud storage. Hence the easy implementation of proposed architecture is reliable for cloud environment.

k) Minimize the response time:

The proposed D1FTC approach designed to implement the two-phase commit protocol to execute transactions faster than other approaches. It allocates the fixed trees depends on the weight of the transactions. It has the shortest path trees to replicate /update the data among cloud virtual machine. So this efficient methodology may minimize the response time for data transactions in cloud.

IX. CONCLUSION AND FEATURE WORK

Cloud provides reliable services to handle analytical data but it faces security and consistency issues when it is offering transactional data management like banking, online reservation and shopping cart, etc. In this paper, the proposed architecture ensures the security in user, service and data storage level with efficient cloud controller and cloud data locker model. The proposed D1FTC approach minimizes the execution time among the virtual machines scatter in the distributed cloud environment. Hence the proposed architecture designed to ensure service level agreement, efficient transaction processing approach, ensure ACID properties, enhanced data storage security for trusted data transactions in cloud environment. The feature work is to implement the proposed architecture in real time cloud environment and do the performance analyze with existing techniques.

X. REFERENCES

[1] Chang Yao, Divyakant Agrawal, Pengfei Chang, Gang Chen,Beng Chin Ooi, WengFai Wong, Meihui Zhang, "DGCC: A New Dependency Graph based Concurrency Control Protocol for Multicore Database Systems", arXiv:1503.03642v1 [cs.DB] 12 Mar 2015.

[2] Jasmina Dizdarevic, Zikrija Avdagic, "The aspects of consistency management of highly-distributed transactional database in a hybrid cloud environment for the energy sector", XI International

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Symposium on Telecommunications (BIHTEL) October 24-26, 2016.

[3] Pornpan Ampaporn and Sethavidh Gertphol, "Performance Measurement of SimpleDB APIs for Different Data Consistency Models", 978-1-4673-7825-3/15/$31.00 ©2015 IEEE.

[4] Peter Bailis, Alan Fekete, Michael J. Franklin, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica, "Feral Concurrency Control:An Empirical Investigation of Modern Application Integrity", Publication rights licensed to ACM, SIGMOD'15, May 31–June 4, 2015.

[5] Álvaro García-Recuero, Sérgio Esteves, Luís Veiga, "Quality-of-Data for Consistency Levels in Geo-replicated Cloud Data Stores", IEEE International Conference on Cloud Computing Technology and Science,2013.

[6] Aleksey Burdakov, Uriy Grigorev, Andrey Ploutenko, Eugene Ttsviashchenko, "Estimation Models for NoSQL Database Consistency Characteristics", 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, 2016.

[7] Jens Kohler and Thomas Specht, "A Performance Comparison Between Parallel and Lazy Fetching in Vertically Distributed Cloud Databases", 978-1-4673-8149-9/15/$31.00 ©2015 IEEE.

[8] Peter Bailis, Alan Fekete, Michael J. Franklin, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica, "Coordination Avoidance in Database Systems", Proceedings of the VLDB Endowment, Vol. 8, No. 3, Copyright 2014 VLDB.

[9] Md. Ashfakul Islam, Susan V. Vrbsky and Mohammad A. Hoque, "Performance Analysis of a Tree-Based Consistency Approach for Cloud Databases", International Conference on Computing, Networking and Communications, Cloud Computing and Networking Symposium, 2012 IEEE.

[10] Stefan Schultea, Christian Janieschb, Srikumar Venugopalc, Ingo Weberd,c, Philipp Hoenischa, "Elastic Business Process Management: State of the Art and Open Challenges for BPM in the Cloud", arXiv:1409.5715v2 [cs.DC] 22 Sep 2014.

[11] Palivela Hemant, Nitin.P.Chawande, Avinash Sonule, Hemant Wani, "DEVELOPMENT OF SERVERS IN CLOUD COMPUTING TO SOLVE ISSUES RELATED TO SECURITY AND BACKUP", Proceedings of IEEE CCIS2011.

[12] Divy Agrawal, Amr El Abbadi, "A Taxonomy of Partitioned Replicated Cloud-based Database Systems", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 38(1), March 2015.

[13] R. ANANDHI, K. CHITRA, "A Challenge in Improving the Consistency of Transactions in Cloud Databases - Scalability", International Journal of Computer Applications (0975 – 8887) Volume 52– No.2, August 2012.

[14] Joarder Mohammad Mustafa Kamal, Manzur Murshed, Rajkumar Buyya, "Workload-Aware Incremental Repartitioning of Shared-Nothing Distributed Databases for Scalable Cloud Applications", IEEE/ACM 7th International Conference on Utility and Cloud Computing, 2014.

[15] Ricardo Jimenez-Peris, Marta Pati˜no-Martinez, Bettina Kemme, "CumuloNimbo: A Cloud Scalable Multi-tier SQL Database", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2015.

[16] Mohd Muntjir, Mohd Asadullah, Shabbir Hassan, Altaf Hussain, "Cloud Database Systems: A Model Conversion in Databases", 2nd International Conference on Information Technology and Electronic Commerce ICITEC 2014.

[17] Thuy D. Nguyen, Mark Gondree, Jean Khosalim, and Cynthia Irvine, "Re-thinking Kernelized MLS Database Architectures in the Context of Cloud-Scale Data Stores", Springer International Publishing Switzerland 2015.

[18] Tahmineh Sanamrad, Lucas Braun, Donald Kossmann and Ramarathnam Venkatesan, "Randomly Partitioned Encryption for Cloud Databases", IFIP International Federation for Information Processing 2014.

[19] Vasily Sidorov, Wee Keong Ng, "Transparent Data Encryption for Data-in-Use and Data-at-Rest in a Cloud-Based Database-as-a-Service Solution", IEEE World Congress on Services, 2015.

[20] M. Thamizhselvan, R. Raghuraman, S. Gershon Manoj, P. Victer Paul, "DATA SECURITY MODEL FOR CLOUD COMPUTING USING V –GRT METHODOLOGY", IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO),2015.

[21] Chuan Yao, Li Xu, and Xinyi Huang, "A Secure Cloud Storage System from Threshold Encryption" 5th International Conference on Intelligent Networking and Collaborative Systems, 2013.

[22] Philip A. Bernstein, Sudipto Das, Bailu Ding, Markus Pilman, "Optimizing Optimistic Concurrency Control for Tree-Structured, Log-Structured Databases", SIGMOD'15, May 31–June 4, 2015.

[23] Md. Ashfakul Islam and Susan V. Vrbsky, "Tree-Based Consistency Approach for Cloud Databases", 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010.

[24] B.Jeevarani, Dr.K.Chitra, "Improved Consistency Model in Cloud Computing Databases ", 2014 IEEE International Conference on Computational Intelligence and Computing Research, 978-1-4799-3975-6/14/$31.00 ©2014 IEEE.

[25] Min Shen, Ajay D. Kshemkalyani, and Ta-yuan Hsu, "Causal Consistency for Geo-Replicated Cloud Storage under Partial Replication", IEEE International Parallel and Distributed Processing Symposium Workshops, IEEE, 2015.

[26] Wenbo Zhu, Murray Woodside, "Tunable Performance & Consistency Tradeoffs for Geographically Replicated Cloud Services (COLOR)", IEEE 2nd International Conference on Cyber Security and Cloud Computing, 2015.

[27] Ayman E. Lotfy , Ahmed I. Saleh , Haitham A. El-Ghareeb , Hesham A. Ali, "A middle layer solution to support ACID properties for NoSQL databases", Journal of King Saud University – Computer and Information Sciences, 2016.

[28] Chao Xie, Chunzhi Su, Cody Littley, Lorenzo Alvisi, Manos Kapritsos and Yang Wang, "High-Performance ACID via Modular Concurrency Control", SOSP'15, October 4–7, 2015, Monterey, CA, 2015.

[29] Zhou Wei, Guillaume Pierre, Chi-Hung Chi, "CloudTPS: Scalable Transactions for Web Applications in the Cloud", IEEE TRANSACTIONS ON SERVICES COMPUTING, SPECIAL ISSUE ON CLOUD COMPUTING, 2011.

[30] Ahmad Waqas, Abdul Waheed Mahessar, Nadeem Mahmood, Zeeshan Bhatti, Mostafa Karbasi, Asadullah Shah, "TRANSACTION MANAGEMENT TECHNIQUES AND PRACTICES IN CURRENT CLOUD COMPUTING ENVIRONMENTS : A SURVEY", International Journal of Database Management Systems (IJDMS) Vol.7, No.1, February 2015.

[31] Marian K. Iskander, Tucker Trainor, Dave W. Wilkinson, Adam J. Lee, and Panos K. Chrysanthis, , "Balancing Performance, Accuracy, and Precision for Secure Cloud Transactions", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 2, FEBRUARY 2014.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[32] Govind Sing, Manmohan and Garima Tiwari, " Cloud Computing: A New Era of IT Opportunity and Challenges", International Journal of Engineering and Management Research, Vol.-3, Issue-4, August 2013.
[33] Palivela Hemant, Nitin.P.Chawande, Avinash Sonule and Hemant Wani, "DEVELOPMENT OF SERVERS IN CLOUD COMPUTING TO SOLVE ISSUES RELATED TO SECURITY AND BACKUP" Proceedings of IEEE CCIS2011, 978-1-61284-204-2/11/$26.00 ©2011 IEEE.
[34] Daniel J. Abadi, "Data Management in the Cloud: Limitations and Opportunities", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2009.
[35] Ms. Shalini Ramanathan, Dr.Savita Goel, Mr.Subramanian Alagumalai, "Comparison of Cloud Databases: Amazon's Simple DB and Google's Big table" International Conference on Recent Trends in Information Systems, 2011.
[36] Francis Gropengießer, Kai-Uwe Sattler, "Transactions a la carte – Implementation and Performance Evaluation of Transactional Support on top of Amazon S3", IEEE International Parallel & Distributed Processing Symposium, 2011.
[37] Balachandra Reddy Kandukuri, Ramakrishna Paturi V, Dr. Atanu Rakshit, "Cloud Security Issues", IEEE International Conference on Services Computing, 2009.
[38] Sebastijan Stoja, Bojan Jelai, Sran Vukmirovi, Darko Capko, Nikola Dalcekovic, "Architecture of Real-Time Database in Cloud Environment for Distributed Systems", Second International Conference on Artificial Intelligence, Modelling and Simulation, 2014.
[39] S. Das, A. El Abbadi, C. Science, and U. C. S. Barbara, "ElasTraS_: An Elastic Transactional Data Store in the Cloud" ,USENIX HotCloud 2, 2009.
[40] S. Das and A. El Abbadi, "G-Store_: A Scalable Data Store for Transactional Multi key Access in the Cloud", SoCC'2010, Indianapolis, Indiana, USA, 2010.
[41] H. T. Vo, C. Chen, and B. C. Ooi, "Towards elastic transactional cloud storage with range query support", Proc. VLDB Endow., vol. 3, no. 1–2, pp. 506–514, Sep. 2010.
[42] G. Decandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo_: Amazon ' s Highly Available Key-value Store", ACM SIGOPS Oper.Syst. Rev., vol. 41, no. 6, 2007.
[43] J. Baker, C. Bond, J. C. Corbett, J. J. Furman, A. Khorlin, J. Larson, L. Jean-michel, Y. Li, A. Lloyd, and V. Yushprakh, "Megastore_: Providing Scalable , Highly Available Storage for Interactive Services", 5th Biennial Conference on Innovative Data Systems Research, 2011.
[44] M. K. Aguilera, A. Merchant, A. Veitch, and C. Karamanolis, "Sinfonia_: A New Paradigm for Building Scalable Distributed Systems", SOSP 2007, Stevenson, Washington, USA, 2007.
[45] Quang Hieu Vu, Maurizio Colombo, Rasool Asal, Ali Sajjad, Fadi Ali El-Moussa and Theo Dimitrakos, "Secure Cloud Storage: A framework for Data Protection as a Service in the multi-cloud environment", IEEE, 2015.
[46] Rohan G. Tiwari, Shamkant B. Navathe, "TOWARDS TRANSACTIONAL DATA MANAGEMENT OVER THE CLOUD", Second International Symposium on Data, Privacy, and E-Commerce, 2010.
[47] Hatem A. Mahmoud, Vaibhav Arora, Faisal Nawab, Divyakant Agrawal, Amr El Abbadi, "MaaT: Effective and scalable coordination of distributed transactions in the cloud", 40th International Conference on Very Large Data Bases, September 1st 5th 2014.
[48] Adam Bates, Ben Mood, Masoud Valafar, and Kevin Butler, "Towards Secure Provenance-Based Access Control in Cloud Environments", CODASPY'13, February 18–20, 2013, San Antonio, Texas, USA. Copyright 2013 ACM.
[49] Marco Serafini, Essam Mansour, Ashraf Aboulnaga, "Accordion: Elastic Scalability for Database Systems Supporting Distributed Transactions", Proceedings of the VLDB Endowment, Vol. 7, No. 12 Copyright 2014.

BIOGRAPHIES

Prof. J. Antony John Prabu is working as an Assistant Professor and pursuing doctor of philosophy in Department of Computer Science, St. Joseph's College,(Autonomous), Tiruchirappalli, Tamil Nadu, India. He received his M.Phil degree from Jamal Mohamed College, Tiruchirappalli. He received his MCA degree from St. Joseph's College, Tiruchirappalli. His areas of interest are Cloud Data Transactions and Distributed Technologies.

Dr. S. Britto Ramesh Kumar is working as an Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has published many research articles in the National/International conferences and journals. His research interests are Cloud Computing, Data Mining, Web Mining, and Mobile Networks.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Color Histogram with Curvelet and CEDD for Content-Based Image Retrieval

Heba A. Elnemr, *Department of Computers and Systems, Electronics Research Institute*

***Abstract*- Content-Based Image Retrieval (CBIR) is one of the most vigorous research areas in the field of pattern recognition and computer vision over the past few years. The accessibility and progressive development of visual and multimedia data, as well as the evolution of the internet, emphasize the necessity to develop retrieval systems that are capable of dealing with a large collection of databases. Many visual features** have been **explored, and it is virtually observed that implementing one kind of features is not efficient in retrieving different types of images. Therefore, in this paper, the author proposes an efficient image retrieval technique that joins color and texture features. The curvelet descriptors that are obtained by using wrapping based discrete curvelet transform are used as texture features. While color features are extracted using quantized RGB color histogram (QCH). Besides, color edge directivity descriptor (CEDD), which joins color and texture features in one histogram is obtained. A multiclass SVM is applied to classify the query images. Four datasets (ALOI, COIL-100, Wang, and Corel-1000) are used to test and assess the proposed system. Improved retrieval results are obtained over CBIR systems based on curvelet descriptors and CEDD individually and jointly. Furthermore, comprehensive experiments have been performed to select the number of histogram bins that achieves an effective and efficient image retrieval. The obtained average precision for the ALOI, COIL-100, Wang and Corel-1000 datasets are 0.996, 998, 0.898 and 0.964, respectively. Also, comparisons with several state-of-the-arts demonstrate the effectiveness of the proposed system in refining the retrieval performance.**

I. INTRODUCTION

Content-based image retrieval (CBIR) is a technique that automatically searches for visually similar images from large scale image databases according to users' requirements. The image retrieval systems based on visual image content have become the center of attention of researchers for more than a decade. The CBIR technique is carried out through two steps; feature extraction and matching policy. In the first step, which considered the most challenging step in CBIR, effective features of each sample image are analyzed and extracted. Most existing general purpose CBIR techniques implement low-level features, such as color, texture, and shape. The set of extracted features is used to build the image signature. While in the second step, the image signatures obtained from the images in the database are compared with that extracted from a query image by a pre-instituted similarity measurement procedure, so that top relevant images in the image database can be returned as the retrieved images [1].

Color content information is one of the most extensively and popularly implemented visual features in CBIR systems. Color feature is comparatively the simplest and most straightforward visual feature for image retrieval. It is also capable of separating images from each other, relatively robust to background complexity and invariant to image size and orientation [2, 3].

The texture is also one of the most commonly utilized low-level visual features in CBIR. Texture features provide spatial and relational information on the intensity distribution over the image [1, 3].

The shape of objects is frequently used as an effective feature for image retrieval because human visual perception can recognize a scene based only on objects shape. Mainly, shape features contain semantic information about an object, thus, the shape description and representation is a very difficult task [3, 4].

During the past years, various techniques have been developed for extracting effective and efficient features. Generally, many developed CBIR techniques are based on extracting a single type of features [4]. However, acceptable retrieval results are hard to be obtained using a single feature type because an image normally

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

comprises various visual characteristics. Consequently, it is mandatory to combine various feature types in such a way as to enhance and emphasize the quality and efficiency of the extracted features so as to obtain an acceptable retrieval performance [5].

This work aims mainly to construct an efficacious CBIR system that is capable of handling large datasets expeditiously. The proposed CBIR system integrates color and texture features through quantized RGB color histogram (QCH), color edge directivity descriptor (CEDD) and curvelet transform.

RGB color model is the most commonly used color model in CBIR, also a color histogram is the simplest and most widespread color feature utilized for image retrieval. QCH denotes the procedure of reducing the number of histogram bins by gathering similar colors into the same bin. Thus, the QCH has a relatively low computational cost, as well as it is invariant to rotation, translation, and scale. However, QCH does not take into consideration spatial distribution or description of color information, this is deemed to be the main weakness of this method. Also, different images could have similar histograms besides minor variation in color due to variations in luminance could produce a considerable change in histogram [6]. Therefore, the author combines QCH with CEDD and curvelet transform descriptors so as to attain a good performance.

CEDD is a powerful low-level feature that joins color and texture information in a single histogram as well as it has comparatively low computational cost, therefore, it is appropriate to be used in huge image databases [7].

On the other hand, curvelet has been widely adopted for image denoising, character recognition, image segmentation, texture analysis as well as image retrieval and it has shown an encouraging performance [3, 8-12]. As curvelet captures more directional features, besides it grabs more accurate texture and directional information and it outperforms wavelet and Gabor filter [3, 12], the author has implemented curvelet descriptors for her proposed retrieval system. Multiclass SVM model is used to accomplish the classification task.

Finally, the author compared the proposed system with others based on CEDD alone, curvelet alone, and CEDD integrated with curvelet. Furthermore, several investigations have been made to choose the appropriate number of histogram bins to attain an efficient and effective retrieval performance. These features prove to be complementary to each other with promising rendering.

The primary contributions of this work can be summarized as follows. (i) The suggested method extracts color and texture information using CEDD and combines it with texture and directional information that are extracted using curvelet transform descriptors. (ii) The proposed CBIR performance is improved using QCH and various experiments are conducted to determine the best number of bins to achieve the best performance. (iii) The proposed CBIR system performance is examined on several forms of large databases including natural, real world and well defined object images.

The paper is organized as follows: Section 2 presents the related work. Section 3 describes the methodology utilized to build the proposed CBIR system while section 4 portrays the depiction of the empirical setup embracing the datasets and the experimental results. Conclusion and future work are discussed in section 5.

II. Related work

Latterly, researchers have proposed many approaches for CBIR using different features such as color, shape, and texture. In this section, the author will discuss several recent literatures that covers some key aspects of CBIR technique.

The work of [12] applied discrete curvelet transform on the Brodatz texture images dataset, then low order statistics is computed from the transformed images. Euclidean distance carried out the similarity measurement in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the suggested CBIR scheme. The results manifest that the proposed curvelet texture feature descriptor performs better than that of Gabor filters in both retrieval efficiency and accuracy.

In [13] a retrieval system that uses local feature descriptors; SIFT and SURF, to generate image signatures invariant to scale and rotation is proposed. Then, BOVW model is created by clustering the local descriptors using K-means technique to build the vocabulary of the K clusters. Finally, the retrieval is accomplished using the SVM classifier model.

Ali et al.[14] designed an image representation scheme based on the histograms of triangles. The proposed method aimed to add spatial information to the inverted index of Bag-of-Features representation. Histograms of triangles are carried out through two levels that are evaluated separately. In the first level, the image is divided into two triangles, while in the second level the image is divided into four triangles. Three different classifiers, Radial Basis Function Neural Networks (RBF-NN), SVM and Deep Belief Networks (DBN), are applied and the overall system performance is evaluated.

Authors of [15] presented an experimental study to investigate and analyze the effect of joining four sampling strategies (SIFT, SURF, Random patch generator and Gauss Random patch generator), with four global feature descriptors (MPEG-7 (Color Layout Descriptor (CLD), Edge Histogram Descriptor (EHD) and Scalable Color Descriptor (SCD)) as well as MPEG-7-like (CEDD), in a Bag-Of-Visual-Word (BOVW) structure. The conducted results demonstrate that the retrieval performance of the proposed descriptors outperforms their performance in their original global form. Besides, they perform better than ordinary SIFT and SURF-based approaches and perform comparably or better against some recent methods.

Malik and Baharudin [16] suggested a CBIR technique that is based on extracting quantized histogram statistical texture features in the compressed domain. The grayscale image is divided into non-overlapping blocks. Next, each block is transformed into a DCT block in the frequency domain. The similarity measurement is achieved using seven distance metrics. The experimental results demonstrate that the Euclidean distance has better performance in both computation and retrieval efficiency.

In the work of [17], a CBIR system with texture and color features succeeded by ant colony optimization feature selection technique is proposed. Wavelet transformation of the sample images is computed and the low-frequency components are used as texture features. Dominant color descriptor, color statistic features, and color histogram features are extracted, in both RGB and HSV domains, as color features. For each feature type, a suitable similarity measure is presented. Ant colony optimization technique is implemented to select the most significant features, among the entire extracted features.

Walia and Pal [18] submitted an image retrieval framework based on combining low-level features. The Color Difference Histogram (CDH) is implemented to extract color and texture features, while Angular Radial Transform (ART) is used to extract shape features. The CDH algorithm is modified in order to improve the overall system performance.

In the work of [19], a CBIR system is designed by combining SURF descriptors with color moments. The similarity strategy is carried out by the KD-tree with the Best Bin First (BBF) search algorithm. Voting Scheme algorithm is finally used to classify and retrieve the matched images from the dataset.

Mukherjee et al. [20] proposed a CBIR system relied on assigning a model of visual words to represent an image patch. Each image patch is represented by a vector that denotes the affinities of the patch for the most significant visual words. To improve the retrieval performance a dissimilarity measure among the pair of image

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

patches is introduced. The dissimilarity measure is made up of two terms: The first one depicts the variation in affinities of the patches that belong to a common set of significant visual words, while the second term penalizes the measure according to the number of visual words that affects only one of the two patches.

III. METHODOLOGY

This paper demonstrates a new CBIR system that relies on extracting texture features as well as color features. The system uses curvelet transform to obtain the spectral domain coefficients that are utilized to compute the texture descriptor of that image while color features are extracted using QCH. Furthermore, CEDD is employed to obtain both color and texture information in a sole histogram. Fig. 1 displays the block diagram of the proposed CBIR system.

A. Quantized RGB Color Histogram

The color histogram is a good portrayal method for describing the color content of an image, it can be obtained by counting the number of occurrences of each color in an image. Pixels in an image are described by three components (typically but not necessarily) in a certain color space, consequently, each pixel is represented as a tuple of three numbers. The RGB color space is the most popular color space used for computer graphics. RGB color histogram investigates each of the RGB-channels separately, this leads to a huge length of the histogram vector (= 256 * 3 for 8-bit RGB image). Thus, color quantization has to be applied in order to produce 3D-color histogram, which is suitable for building efficient indexes for large image databases as well as has an acceptable computational cost. In color quantization procedure, the number of colors used to represent an image is reduced, and each color component is quantized into a number of 'bins'. Since color components (R, G, and B) are equally important, each component is quantized into the same number of bins. In this work, the QCH is implemented and a various number of bins are tested in order to find the best quantization level.

B. CEDD histogram

CEDD is a low-level feature that incorporates color and texture information in a single histogram. One of the greatest significant characteristic of CEDD, the low computational power required for its extraction in comparison with that needed for most MPEG-7 descriptors [7, 21]. CEDD is linked with a texture unit to extract textural information besides a color unit to extract color information.

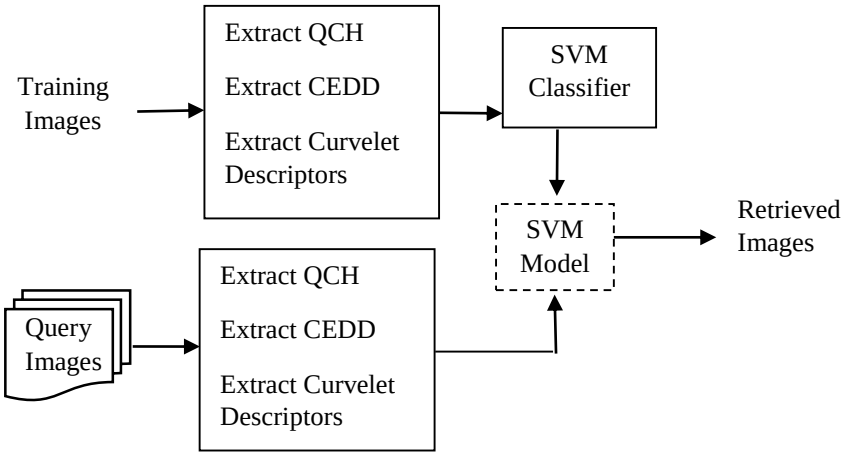


Figure 1. A block diagram of the proposed CBIR system.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In order to obtain the CEDD histogram, each image is divided into 1600 image blocks, and then each block is processed through the following algorithms. In the texture unit, the image block is parted into 4 regions, sub-blocks, the mean value of the luminosity of the pixels enclosed within each sub-block is considered to be the value of this sub-block. The luminosity values are computed within the YIQ color space. Afterward, each block is filtered utilizing 5 digital filters that were suggested in the MPEG-7 Edge Histogram Descriptor [7, 21].

On the other hand, in the color unit, a set of fuzzy rules are implemented to obtain the color information. Each image block is transferred into HSV color space. Next, Fuzzy-Linking histogram technique [22] is applied. The fuzzy system creates a 24-bins histogram.

The CEDD structure incorporates six texture regions and each of which holds 24 color regions. Accordingly, the final histogram consists of 6 × 24 = 144 bins, conforming to the total 144 color regions within the six texture regions. Finally, the histogram is normalized and quantized so as each bin is represented by 3 bits. Hence, its total size is limited to be 144 x 3 = 432 bits or 54 bytes per image. This small size of the descriptor is considered to be another advantage of CEDD.

C. Curvelet Transform

Curvelet transform is a multi-scale transform designed to represent images at various scales and various angles, this transform is established by Donoho and Duncan in 1999 [23]. In this transform, the input image is initially decomposed into a set of subbands and each subband is partitioned into blocks that are analyzed by ridgelet transform. The ridgelet transform is realized using the Radon transform and the 1-D wavelet transform [24]. To avoid blocking effect, the spatial partitioning process involves overlapping of windows that lead to a large amount of redundancy. As well, it is very time-consuming, which makes it inappropriate for texture feature analysis in a large database. Therefore, Candes et al. [25] developed fast discrete curvelet transform that is based on the wrapping of Fourier samples. This transform is simpler, faster, less redundant, and has less computational complexity since it applies fast Fourier transform instead of the complex ridgelet transform.

For a 2-D input image of size M x N, the curvelet transform based on the wrapping of Fourier samples generates a set of curvelet coefficients indexed by a scale j, an orientation l, and two spatial location parameters (k1, k2). These coefficients are defined as follows [25]

$$C_{j,l}(k_1,k_2)=\sum_{\substack{0\le m<M\\0\le n<N}} f(m,n)\,\varphi^{D}_{j,l,k_1,k_2}(m,n) \tag{1}$$

where f(m, n) is the Cartesian array of the input image and $\varphi^{D}_{j,l,k_1,k_2}(m,n)$ is a digital Curvelet waveform. These coefficients are then used to form the curvelet texture descriptors by implementing statistical operations.

Curvelet Texture Features Extraction

After the curvelet coefficients in each sub-bands are created and stored, curvelet statistical features, i.e. mean and standard deviation, of the coefficients corresponding to each sub-bands, are computed. These features proved to be capable of describing curvelet sub-bands [12, 26]. In general, these mean and standard deviation are then used as texture descriptors of the image. Hence, for each curvelet sub-band, the author obtains two texture features. If n curvelets sub-bands are applied for the transform, 2n dimensional texture feature vector is obtained for each image.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 1
CURVELET TRANSFORM (4 LEVEL DECOMPOSITION)

Scale	1	2	3	4
No. sub-band	1	16	32	1
Sub-bands considered for feature calculation	1	8	16	1

In this work, 4 level curvelet transform is implemented to decompose the input images. Based on this analysis, 50 (=1+16+32+1) curvelet coefficients are computed. However, [12] presents that curvelet at an angle θ generates the same coefficients as curvelet at an angle (θ + π). Thus, due to this symmetry, half of the sub-bands at scale 2 and 3 may be rejected. Accordingly, 26 (=1+8+16+1) sub-bands are maintained, producing a 52 dimension feature vector for each image in the database. Table 1 illustrates the sub-bands distribution at each transform level.

The mean of a sub-band at scale s and orientation θ is stated as:

$$\mu_{s,\theta} = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} C_{s,\theta}(i, j) \tag{2}$$

while the standard deviation of a sub-band at scale s and orientation θ is expressed as:

$$\sigma_{s,\theta} = \sqrt{\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (C_{s,\theta}(i, j) - \mu_{s,\theta})^2} \tag{3}$$

where Cs,θ(i, j) represents the curvelet coefficient at scale s, orientation θ and location (i, j).

IV. RESULTS AND ANALYSIS

A. Image Datasets

The proposed system was assessed using four different standard datasets; Amsterdam Library of Object Images (ALOI) dataset [27], Columbia object image library (COIL-100) dataset [28] and two subsets of the Corel image database [29].

ALOI image dataset provides one-thousand small objects recorded under various imaging conditions (viewing angle, illumination angle, and illumination color). More than a hundred images of each object were registered. In this work, 102 objects are randomly selected with 72 different viewpoints each. COIL-100 is a well-known standard color image database that includes 100 objects with 72 views acquired by rotating each object around the vertical axis. Two subsets of Corel database are utilized, each of which consists of 10 irrelevant classes and each contains 100 images of the Corel stock photo database. The first subset is the Wang database that contains: Africa, Beach, Building, Bus, Dinosaur, Elephant, Flower, Horse, Mountain and Food classes. While the second one is Corel-1000 that comprises; Dinosaur, Cyber, Horse, Bonsai, Texture, Fitness, Dishes, Antiques, Elephant and Easter egg groups. Fig. 2, Fig. 3, Fig. 4 and Fig. 5 illustrate samples of the utilized datasets.

These databases are selected to represent two different types of CBIR chores: The ALOI and COIL-100 datasets characterize retrieval task that involves depicting more clearly defined objects with various viewing angles while Wang and Corel-1000 databases signify retrieval task with real world arbitrary photographs.

B. Implementation Details

All assessments were accomplished on a Lenovo laptop with Intel Core i7, 2.20 GHz processor, 8GB RAM, and Windows 10 Home Ultimate 64-bit as an operating system. The system was implemented in Matlab R2013b.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

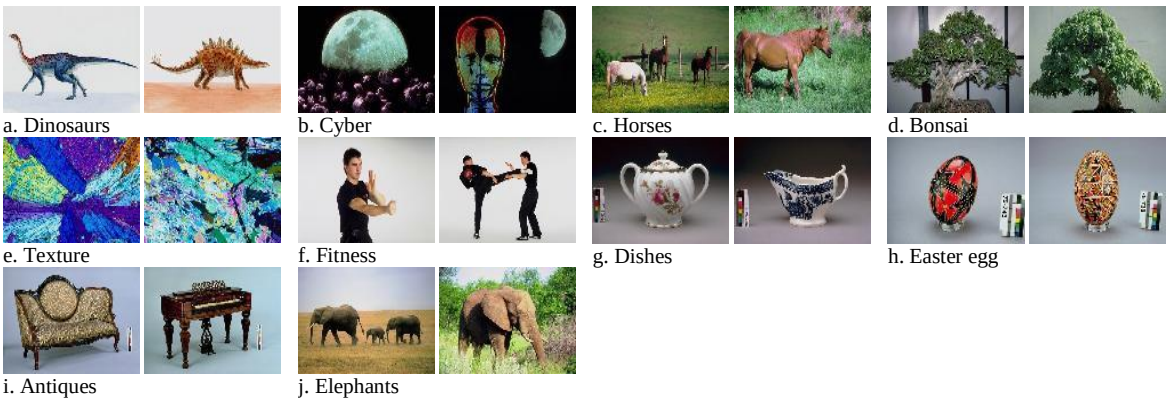
In this work, two validation techniques are utilized, viz. repeated holdout validation and k-fold cross-validation. The repeated holdout method splits randomly the data into two disjoint subsets; training set and test set and repeats this process with different subsamples. On the other hand, the k-fold validation generates a k subsets of equal size. The system is trained with k – 1 subsets and the remaining one forms the test set. This procedure is repeated k times. The holdout is simpler and needs less computation, however, there are overlapping test sets. On the contrary, k-fold cross-validation has the advantage that there are non-overlapping test sets. All samples in the dataset are ultimately used for both training and testing. However, it is computationally expensive.

Holdout validation is achieved by randomly choosing 90% of a dataset images for training and the remaining 10% of the images are reserved for testing. This validation procedure is repeated five times and the average performance is computed. Besides, ten-fold cross-validation is employed to evaluate the system performance.

To evaluate the proposed retrieval system, 4-level curvelet transform is applied. Besides, different quantization levels are used to compute the color histogram (CH) so as to obtain the optimum number of bins. The tests are carried out on 5 quantization levels; 9, 16, 25, 64 and 100 bins for each of the RGB channels.

Figure 2. Sample images from ALOI database.

Figure 3 . Sample images from COIL-100 database.



a. Dinosaurs b. Cyber c. Horses d. Bonsai
e. Texture f. Fitness g. Dishes h. Easter egg
i. Antiques j. Elephants

Figure 4. Sample images from Corel-1000 database.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

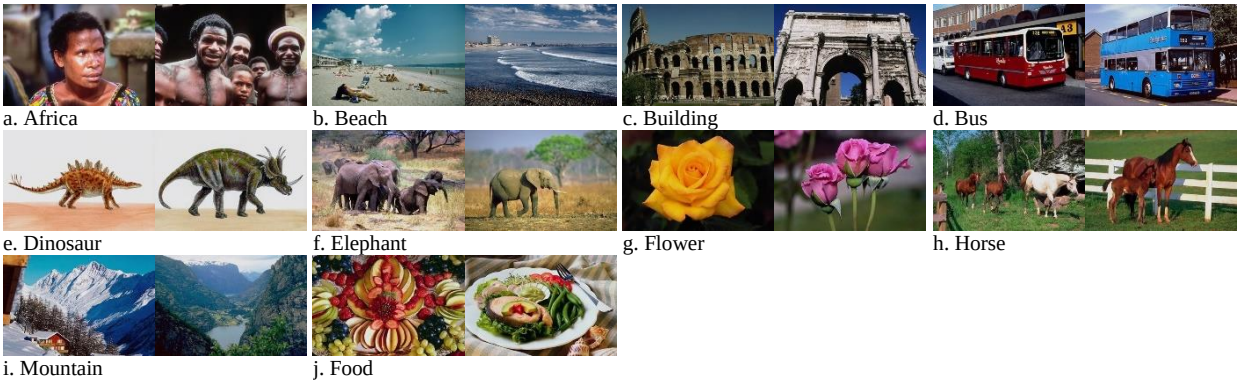


Figure 5. Sample images from Wang database.

Recall and precision metrics are utilized to measure the performance of the proposed CBIR system. The two metrics are defined as follows:

$$\mathrm{Pr}ecision = \frac{Number\ \ of\ \ relevant\ \ images\ \ retrieved}{Total\ \ number\ \ of\ \ images\ \ retrieved} \tag{4}$$

$$\mathrm{Re}call = \frac{Number\ \ of\ relevant\ \ images\ \ retrieval}{Total\ \ number\ \ of\ \ relevant\ \ images\ \ in\ \ the\ \ database} \tag{5}$$

Furthermore, Precision-Recall curve (PRC) is used to assess the effectiveness of the proposed image retrieval system.

C. Experimental results and discussion

In this section, the author presents and debates the results of the experimental evaluation of the proposed retrieval system. To estimate the efficiency of all implemented features (curvelet descriptors, quantized RGB CH (QCH) and CEDD) on the given databases, the author extracted these features from the images and executed experiments to test the effect of curvelet descriptors and CEDD features individually, the combination of both as well as the combination of them with the QCH. In addition, the author scrutinized the effect of using different quantization levels (five quantization levels are studied; 9, 16, 25, 64, 100). For all experiments, the author reports the average precision and recall ratios for the holdout and K-fold validation methods. Moreover, the results are benchmarked with previous works that use the same databases utilized in this work. Also, the author has compared the results of her previous work [30] with that of this work.

In this work, each image is represented by a feature vector of size

$$\begin{aligned} &feature\ \ vector\ \ size = N_{curvelet} + N_{CEDD} + N_{CH} \\ &N_{CH} = i \times 3 \end{aligned} \tag{6}$$

where $N_{curvelet}$ is the dimension of curvelet descriptor (52 for 4-level decomposition), N_{CEDD} is the dimension of CEDD vector (144), N_{CH} is the total number of bins of the 3D-color histogram and i is the number of bins for each RGB channel (since RGB channels are equally quantized). Table 2 depicts the dimensions of the extracted feature vectors for the different realized methods.

Fig. 6 and Fig. 7 represent the experiments conducted on ALOI dataset using different retrieval methods. The results indicate that joining CEDD and curvelet improves the retrieval system significantly and combining them

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

with color histogram descriptors further enhances the retrieval performance. The optimal precision using K-fold cross-validation (0.995) and using holdout validation (0.996) is attained when the quantization level is 16 and 100 for each RGB channel, whereas it is near optimal at quantization levels 9, 25 and 64, with tiny variations. While the best recall (0.994) is obtained with 25 quantization levels using holdout validation as well as 100 quantization levels using K-fold validation. The proposed system achieved a good performance during both K-fold and holdout validations. This demonstrates the efficiency of the proposed retrieval system.

Fig. 8 and Fig. 9 summarize the experimental results of all implemented descriptors on the COIL-100 dataset. The results reported in Fig. 8 indicate that the best precision ratio (0.998) is reached by integrating CEDD, curvelet and RGB color histogram with 25 bins using holdout validation and with 9 and 25 bins using K-fold validation. As well, it is nearby to optimum (0.997), with an insignificant difference, when using 16 bins for both methods of validation. Additionally, Fig. 9 illustrate that the best recall ratio (0.998) is attained by joining CEDD, curvelet and RGB color histogram with 9 and 25 bins employing K-fold validation, and it is near to optimal (0.997) using 16, 25 and 64 bins using holdout validation as well as 64 bins using K-fold validation. It is also worth to be noted that recall close to optimum (0.997) when joining CEDD and curvelet descriptors for the holdout validation.

Fig. 10 and Fig. 11 report the results obtained by employing the proposed retrieval techniques on the Corel-1000 dataset. The author realized from these figures that integrating CEDD and curvelet descriptors enhances the precision ratio for both validation methods, the precision reaches its maximum value (0.954) using K-fold validation. Moreover, merging these descriptors with RGB color histogram with 25 bins further improves the precision value (0.964), and it is close to optimal (0.955) with 16 bins using holdout validation as well as (0.951) with 9 bins using K-fold validation. Also, utilizing 16 and 25 bins yield to almost the best recall ratio (0.948 and 0.943 using K-fold and holdout validation, respectively).

Fig. 12 and Fig. 13 present the comparison results of the Wang dataset using the proposed retrieval methods. From the results, the author recognized that the optimal precision ratio is reached utilizing CEDD, curvelet and RGB color histogram with 16 bins (0.898) using holdout validation and 9 bins (0.892) using K-fold validation with a tiny difference (0.006). Furthermore, recall ratio achieves the best results when employing CEDD, curvelet and RGB color histogram with 16 bins (0.874 using holdout validation) and 9 bins (0.877 using K-fold validation).

It can be noticed from the results that CBIR using the CEDD has a better performance than using curvelet this is because CEDD has color and texture features. Moreover, joining both descriptors enhances the performance significantly. Besides, integrating QCH descriptor improves the retrieval efficiency further and this improvement varies as the quantization levels vary. The author can also perceive that the use of greater quantization levels does not necessarily lead to a better precision. Contrarily, it leads to a much less efficient search. In CBIR, the retrieval effectiveness is essential than slight precision gain. Small retrieval performance enhancement at the cost of much higher dimension will reduce the entire system efficiency. Thus, for retrieval efficiency, the author recommends that CEDD, curvelet and RGB color histogram obtained from either 9, 16 or 25 quantization levels are appropriate to accomplish the retrieval tasks for all stated datasets.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 2
THE FEATURE VECTOR SIZE FOR THE IMPLEMENTED RETRIEVAL METHODS

Method	Feature vector size
CEDD	144
Curvelet	52
CEDD + Curvelet	196
CEDD + Curvelet + QCH with 9 bins	223
CEDD + Curvelet + QCH with 16 bins	244
CEDD + Curvelet + QCH with 25 bins	271
CEDD + Curvelet + QCH with 64 bins	388
CEDD + Curvelet + QCH with 100 bins	496

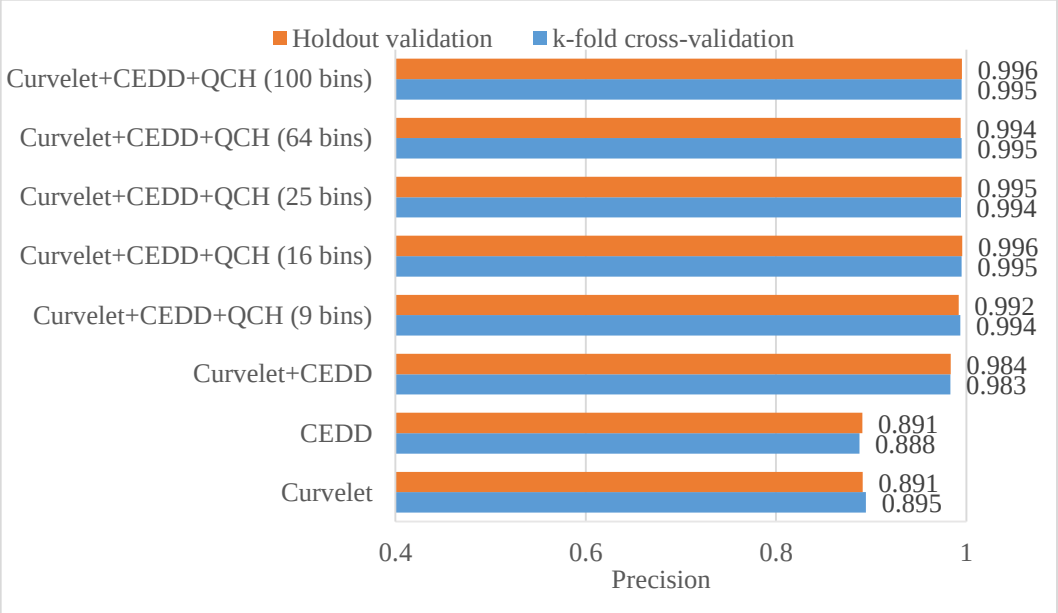


Figure 6. Average precision for different methods on ALOI dataset

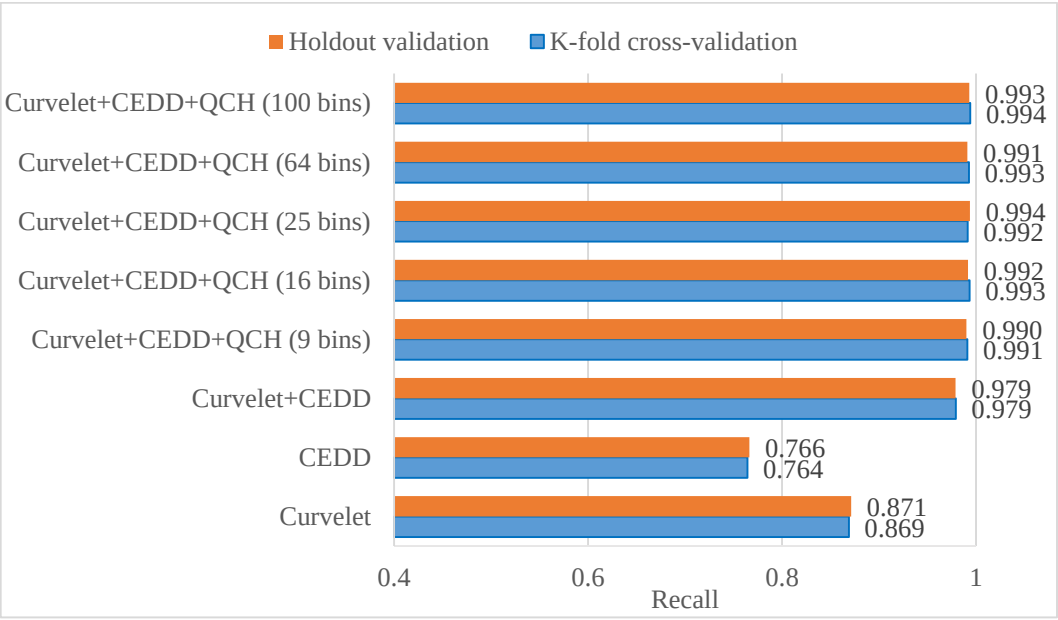


Figure 7. Average recall for different methods on ALOI dataset

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

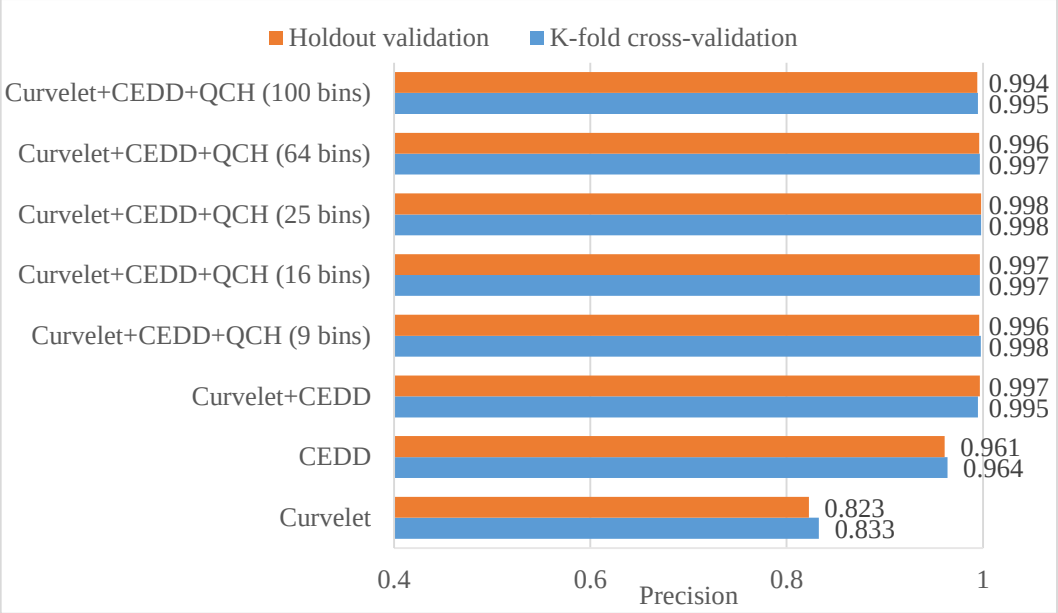


Figure 8. Average precision for different methods on COIL-100 dataset

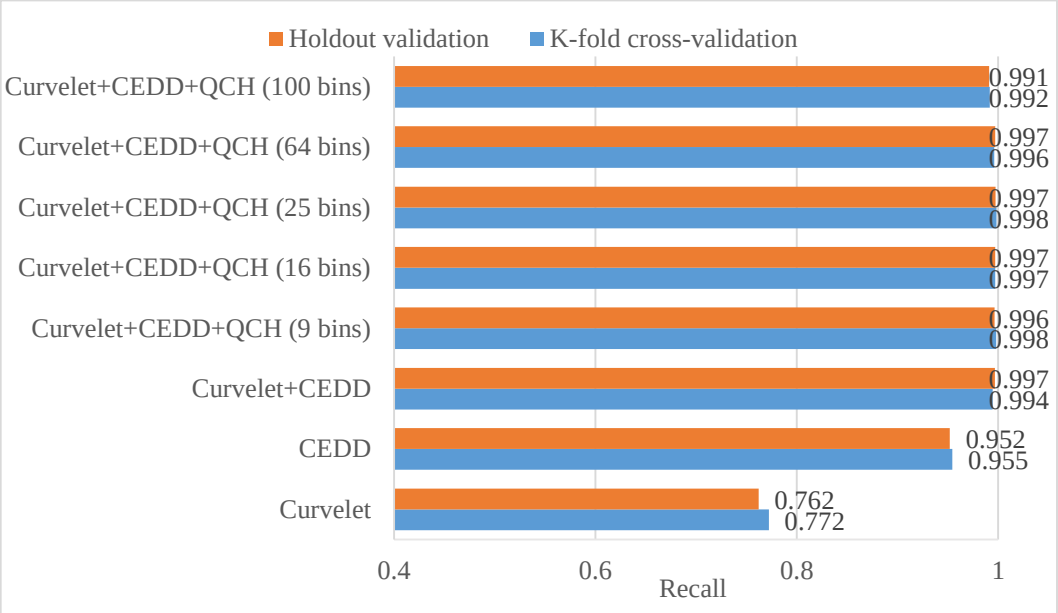


Figure 9. Average recall for different methods on COIL-100 dataset

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

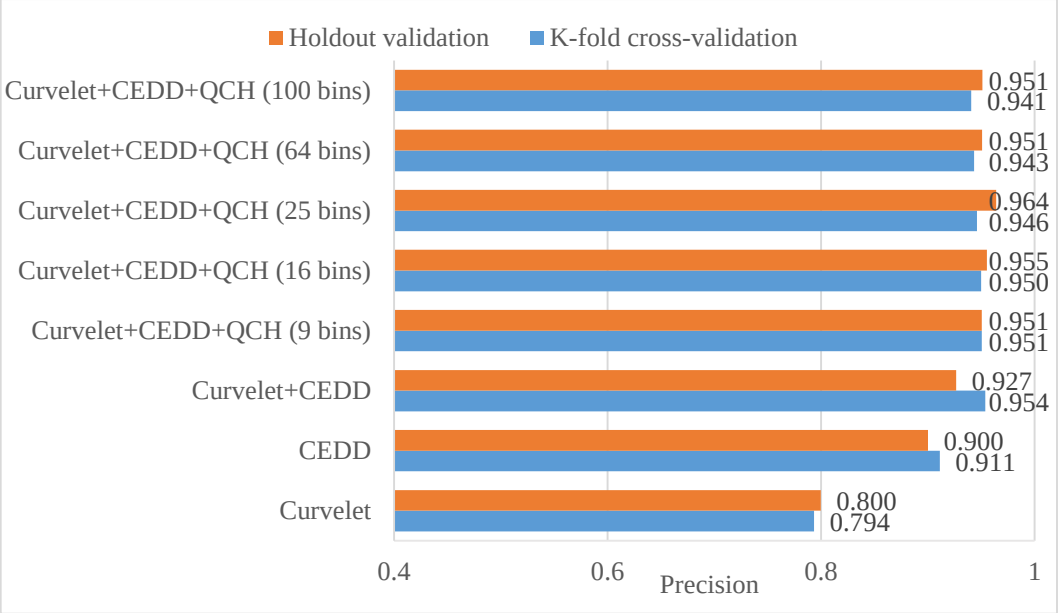


Figure 10. Average precision for different methods on Corel-1000 dataset

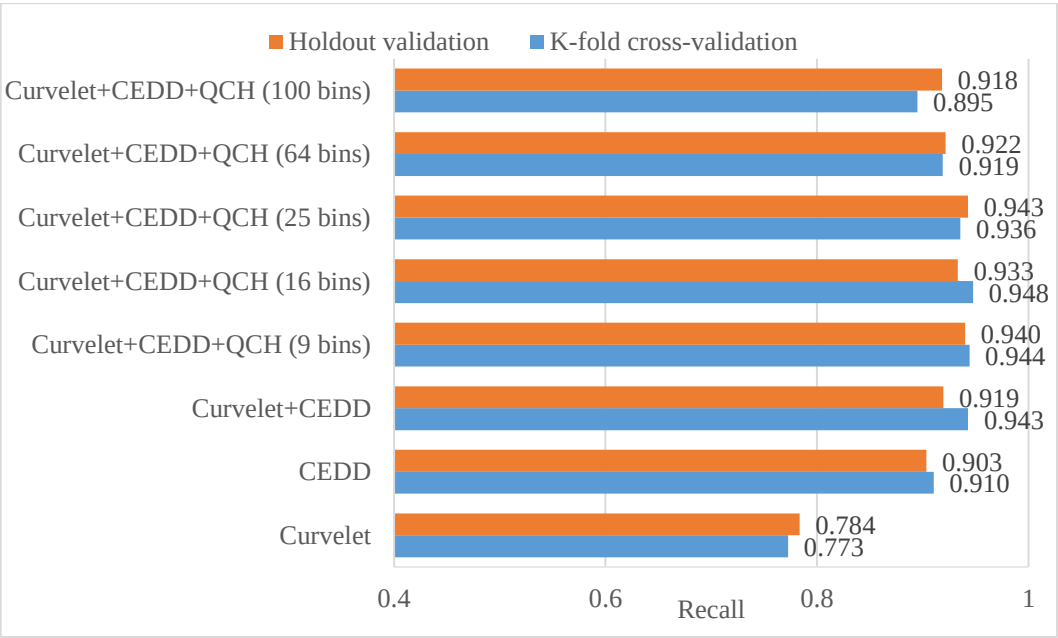


Figure 11. Average recall for different methods on Corel-1000 dataset

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

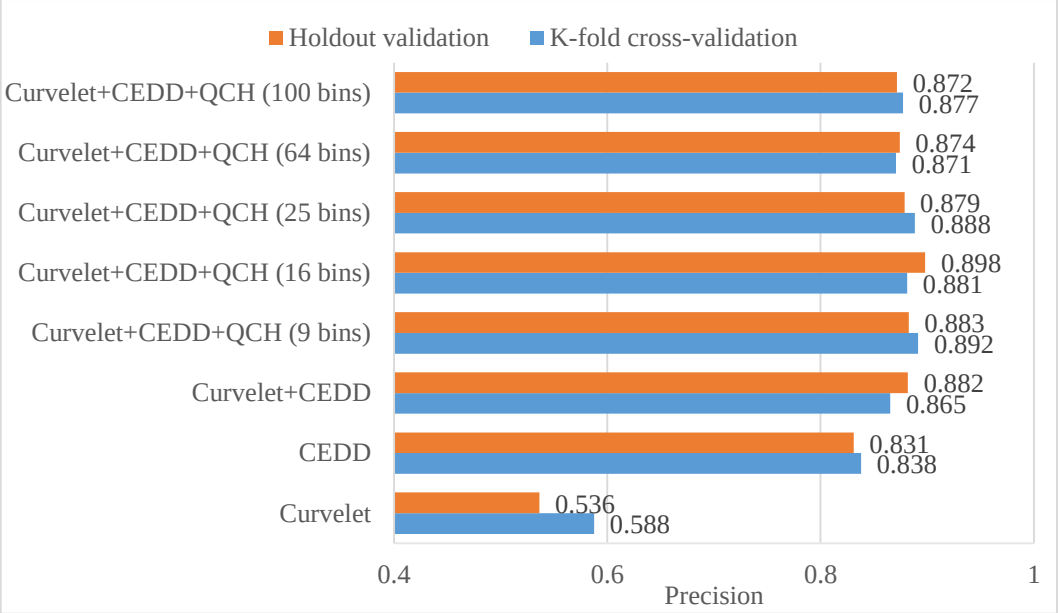


Figure 12. Average precision for different methods on Wang dataset

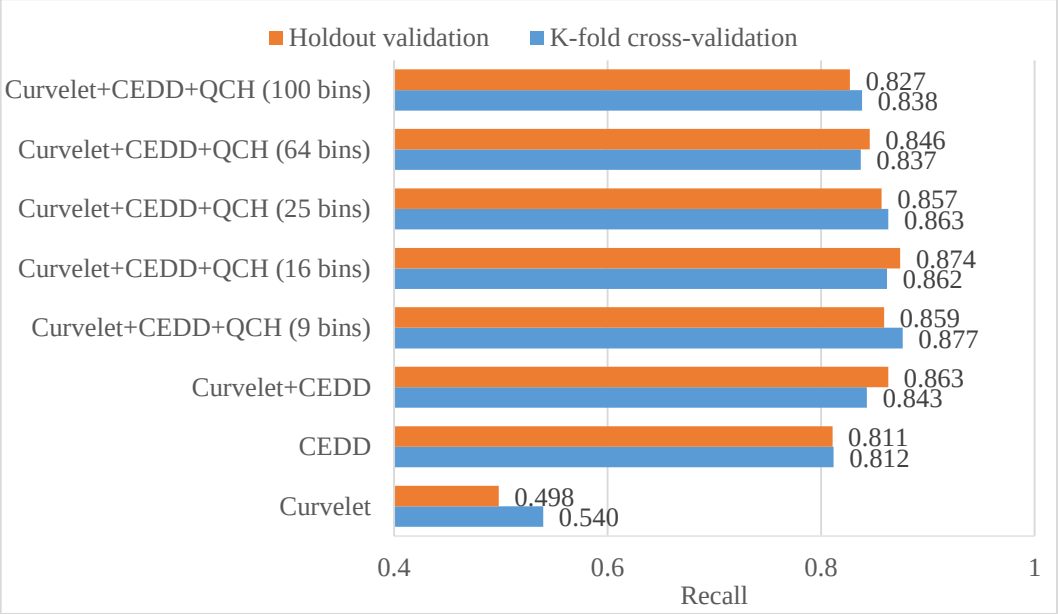


Figure 13. Average recall for different methods on Wang dataset

Moreover, the author has realized from the results that different features perform differently on the various datasets. Fig. 14 and Fig. 15 present the average precision of applying CEDD, curvelet and RGB color histogram with 9, 16 and 25 bins descriptors over all databases using K-fold and holdout validations, respectively. It is obvious that the system performs very well on ALOI and COIL-100 datasets compared to the Wang and Corel-1000 datasets, this is mainly because of the consistent background in all images in case of ALOI and COIL-100 datasets.

Furthermore, the author compared her retrieval method with other existing techniques. Table 3, table 4, table 5 and table 6 report comparisons between the proposed method and a group of other techniques on the explored

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

datasets. The performance of the proposed system, evaluated in terms of precision, achieved good results and the proposed model is competent with all the compared models.

The results obtained on ALOI and Corel-1000 datasets denote that the proposed system significantly outperforms previously addressed methods. However, for COIL-100 dataset, the method suggested in [35] shows comparable retrieval precision. Nevertheless, this method examines only 10 objects that were selected randomly from the whole dataset.

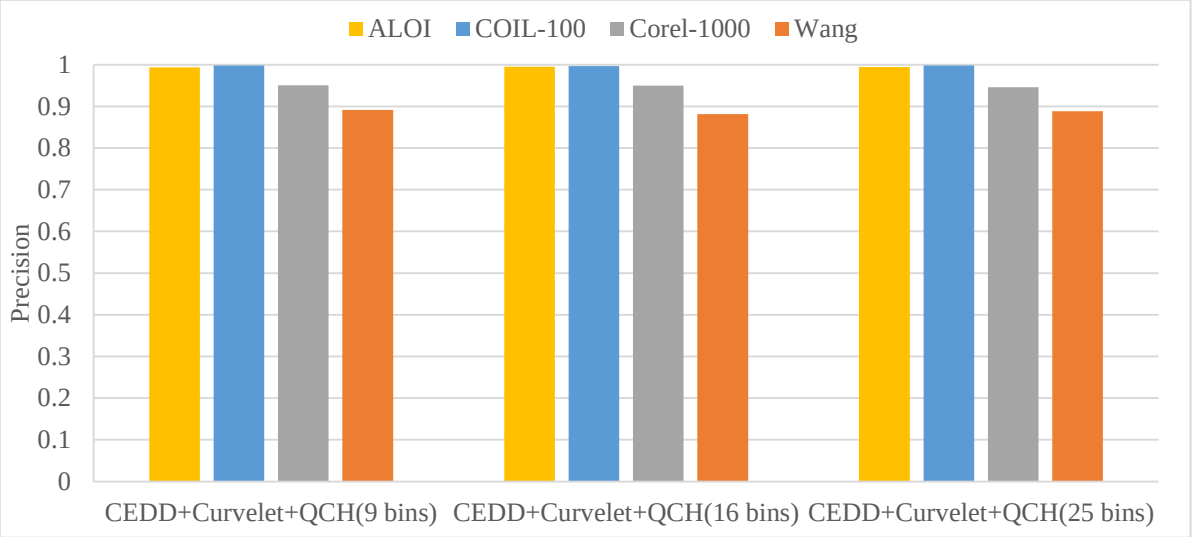


Figure 144. Average precision using K-fold validation different extracted features over each database

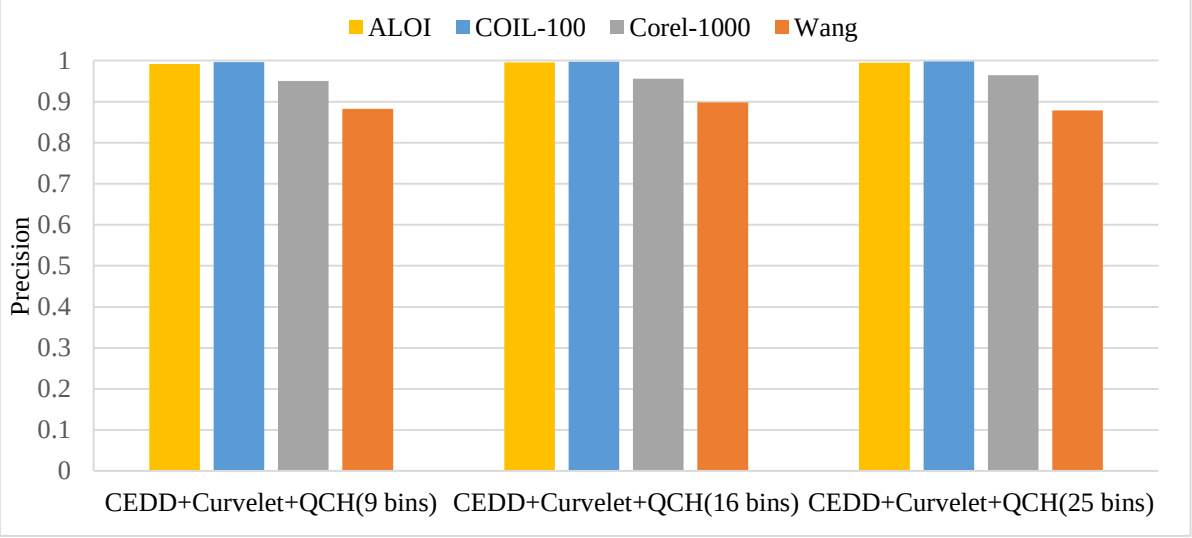


Figure 155. Average precision using holdout validation for different extracted features over each database

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 3
COMPARISON WITH OTHER METHODS ON COIL-100 DATASET

Method	Precision
The proposed method (with 25 bins)	0.998
Kavya and Shashirekha [31] (10 random objects were only considered)	0.86
Velmurugan and Baboo [19] (15 random objects were only considered)	0.88
Mukherjee et al. [20] (10 random objects were only considered)	0.86
Bahri and Zouaki [32] (15 random objects were only considered)	0.78
Elnemr [30]	0.93
Khatabi et al. [35] (10 random objects were only considered)	0.9985

TABLE 4
COMPARISON WITH OTHER METHODS ON ALOI DATASET

Method	Precision
The proposed method (with 25 bins)	0.995
Alkhawlani, Elmogy and Elbakry [13] (10 random objects were only considered)	0.88

TABLE 5
COMPARISON WITH OTHER METHODS ON WANG DATASET

Method	Precision
The proposed method (with 16 bins)	0.898
Ali et al. [14]	0.877
Iakovidou et al. [15]	0.82
Rashno, Sadri and SadeghianNejad [17]	0.63
Mehmood et al. [33]	0.84
Vilvanathan and Rangaswamy [34]	0.75

TABLE 6
COMPARISON WITH OTHER METHODS ON COREL-1000 DATASET

Method	Precision
The proposed method (with 4 bins)	0.95
Elnemr [30]	0.88

V. CONCLUSION

This paper proposes a new CBIR technique that is based on integrating CEDD, curvelet and QCH descriptors. The classification stage is performed using a multiclass SVM. Generally, the precision of a CBIR system decreases as the number and variety of images increases in the dataset. Thus, the author assessed her proposed retrieval technique on benchmark databases from various domains such as to cover a wide range of different CBIR applications. The performance analysis is evaluated by computing the precision and recall as metrics. K-fold cross-validation and holdout validation are used to validate the results of implementing the various investigated descriptors as well as the different quantization levels. The experimental results are analyzed on the basis of comparing the retrieval performance of CEDD and curvelet individually and jointly as well as integrating CEDD, curvelet, and different QCH bins. The results indicate that merging CEDD and curvelet descriptors enhance the retrieval performance significantly. Furthermore, integrating them with QCH improves the performance more. The author also concludes that combining CEDD, curvelet and RGB color histogram with 9, 16 or 25 bins descriptors outperformed other examined descriptors regarding efficiency and scalability. Additionally, comparisons with existing CBIR techniques illustrate the effectiveness and efficiency of the proposed method.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Thus, the author hereby demonstrates the prospect of creating a better CBIR system with more significant feature sets.

In future, the employed datasets can be increased and new classes can be added to design a generalized and efficient retrieval system. Furthermore, advanced techniques like deep learning may be used to develop an efficient system for image retrieval and annotation.

REFERENCES

[1] M. Yasmin, S. Mohsin, and M. Sharif, "Intelligent image retrieval techniques: A survey," *Journal of Applied Research and Technology*, vol. 12, no. 1, pp. 87-103, Feb. 2014.
[2] K. Jenni, S. Mandala, and M. Sunar, "Content based image retrieval using colour strings comparison," *Procedia Computer Science*, vol. 50, pp. 374-379, 2015.
[3] H. Elnemr, N. Zayed, and M. Fakhreldein, "Feature extraction techniques: fundamental concepts and survey," in Handbook of Research on Emerging Perspectives in Intelligent Pattern Recognition, Analysis, and Image Processing, N. Kamila, Ed. IGI Global, 2016, pp. 264-294.
[4] S. Brandt, J. Laaksonen, and E. Oja, "Statistical shape features for content-based image retrieval," *Journal of Mathematical Imaging and Vision*, vol. 17, no. 2, pp. 187-198, Sep. 2002.
[5] S. Gandhani, and N. Singhal, "Content based image retrieval: survey and comparison of CBIR system based on combined features," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 10, pp. 155-162, Oct. 2015.
[6] K. Meskaldji, S. Boucherkha, and S. Chikhi, "Color quantization and its impact on the color histogram based image retrieval accuracy," in *NDT '09 First International Conf. on Networked Digital Technologies*, 2009, pp. 515-517.
[7] S. Chatzichristofis, and Y. Boutalis, "CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in Computer Vision Systems, A. Gasteratos, M. Vincze and J. Tsotsos, Ed. Berlin Heidelberg: Springer-Verlag, 2008, pp. 312-322.
[8] J. Starck, E. Candès and D. Donoho, "The Curvelet transform for image denoising," *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 670-684, Jun 2002.
[9] M. Angshul, "Bangla basic character recognition using digital Curvelet transform," *Journal of Pattern Recognition Research*, vol. 2, no. 1, pp. 17-26, 2007.
[10] G. Joutel, V. Eglin, S. Bres, and H. Emptoz, "Curvelet based feature extraction of handwritten shapes for ancient manuscripts classification," in *Proc. of SPIE-IS&T Electronic Imaging*, Jan 2007, pp. 1-12.
[11] Rajakumar, and Muttan, "A framework for MRI image retrieval using Curvelet transform and Euclidean distance," *Journal of Computer Science*, vol. 9, no. 3, pp. 285-290, 2013.
[12] I. J. Sumana, M. Islam, D. Zhang, and G. Lu, "Content based image retrieval using curvelet transform," in *IEEE 10th workshop on Multimedia signal processing*, Australia, Oct. 2008, pp. 11-16.
[13] M. Alkhawlani, M. Elmogy and H. Elbakry, "Content-based image retrieval using local features descriptors and Bag-of-Visual Words," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 6, no. 9, pp. 212-219, 2015.
[14] N. Ali, K. B. Bajwa, R. Sablatnig, and Z. Mehmood, "Image retrieval by addition of spatial information based on histograms of triangular regions," *Computers and Electrical Engineering*, vol. 54, pp. 539–550, Aug. 2016.
[15] C.Iakovidou, N. Anagnostopoulos, A. Kapoutsis, Y. Boutalis, M. Lux, and S. Chatzichristofis, "Localizing global descriptors for content-based image retrieval," *EURASIP Journal on Advances in Signal Processing*, vol. 80, pp. 1-20, Dec. 2015.
[16] F. Malik, and B. Baharudin, "Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain," *Journal of King Saud University – Computer and Information Sciences*, vol. 25, no. 2, pp. 207–218, July 2013.
[17] A. Rashno, S. Sadri, and H. SadeghianNejad, "An efficient content-based image retrieval with ant colony optimization feature selection schema based on wavelet and color features," in *International Symposium on Artificial Intelligence and Signal Processing (AISP)*, Iran, March 2015, pp. 59-64.
[18] E. Walia, and A. Pal, "Fusion framework for effective color image retrieval," *Journal of Visual Communication and Image Representation*, vol. 25, no. 6, pp. 1335–1348, Aug. 2014.
[19] K. Velmurugan, and S. Baboo, "Content-based image retrieval using SURF and color moments," *Global Journal of Computer Science and Technology*, vol. 11, no. 10, pp. 1-5, May 2011.
[20] A. Mukherjee, S. Chakraborty, J. Sil, and A. Chowdhury, "A novel visual word assignment model for Content-Based Image Retrieval," in *Proceedings of International Conference on Computer Vision and Image Processing, Advances in Intelligent Systems and Computing*, Springer, Singapore, 2017, pp. 79-87.
[21] K. Jayanthi and M. Karthikeyan, "Color and edge directive descriptor feature extraction technique for content based image retrieval system," *Middle-East Journal of Scientific Research*, vol. 23, no. 8, pp. 1590-1597, 2015.
[22] S. Chatzichristofis, and Y. Boutalis, "A hybrid scheme for fast and accurate image retrieval based on color descriptors," in *IASTED International Conference on Artificial Intelligence and Soft Computing (ASC 2007)*, Spain, Aug. 2007, pp. 280-285.
[23] E. Candès, and D. Donoho, "Curvelets - a surprisingly effective nonadaptive representation for objects with edges," Department of Statistics, Stanford University, Report Number 1999-28, Nov. 1999.
[24] J. Starck, E. Candès and D. Donoho, "The Curvelet transform for image denoising," *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 670-684, Aug. 2002.
[25] E. Candès, L. Demanet, D. Donoho, and L. Ying, "Fast discrete Curvelet transforms,"*Multiscale Modeling and Simulation*, vol. 5, no. 3, pp. 861-899, March 2006.
[26] S. Arivazhagan, L. Ganesan, and T. Kumar, "Texture classification using Curvelet statistical and co-occurrence features," presented at the 18th International Conference on Pattern Recognition (ICPR'06), Aug. 20-24, 2006, China.
[27] J. Geusebroek, G. Burghouts, and A. Smeulders, "The Amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103-112, 2005.
[28] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Department of Computer Science, Columbia University, Technical Report CUCS-006-96, March 1996.
[29] J. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries*," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 23, no. 9, pp. 947-963, Sep. 2001.
[30] H. Elnemr, "Combining SURF and MSER along with color features for image retrieval system based on bag of visual words," *Journal of Computer Science*, vol. 12, no. 4, pp. 213-222, 2016.

[31] J. Kavya, and H. Shashirekha, "A Novel approach for image retrieval using combination of features," *International Journal of Computer Technology & Applications*, vol. 6, no. 2, pp. 323-327, Jan. 2015.
[32] A. Bahri, and H. Zouaki, "A SURF-Color moments for images retrieval based on bag-of features," *European Journal of Computer Science and Information Technology*, vol. 1, no. 1, pp. 11-22, Jun 2013.
[33] Z. Mehmood, S. Anwar, N. Ali, H. Habib, and M. Rashid, "A novel image retrieval based on a combination of local and global histograms of visual words," *Mathematical Problems in Engineering*, vol. 2016, pp. 1-12, Jun 2016.
[34] K. Vilvanathan, and R. Rangaswamy, "Bi-Level classification of color indexed image histograms for content based image retrieval," *Journal of Computer Science*, vol. 9, no. 3, pp. 343-349, 2013.
[35] Khatabi, A. Tmiri and A. Serhir, "An efficient method of improving image retrieval using combined global and local features," Advances in Ubiquitous Networking 2. Lecture Notes in Electrical Engineering, Springer, Singapore, 2017, pp. 431-443.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A survey on multi-objective task scheduling algorithm in cloud environment

Tanvi Gupta **Dr.SS.Handa** **Dr. Supriya Panda**

Research Scholar **Professor** **Professor**

Manav Rachna International Institute of Research and Studies SRM University Manav Rachna International Institute of Research and Studies

Abstract: *Cloud computing is one of the important subject now- a- days in which services are given to the users by cloud provider. So, according to the use of the services users will pay to the providers. Resource allocation and task scheduling are important to manage the task in cloud environment for load balancing. Task scheduling is an important step to improve the overall performance of the cloud computing. Task scheduling is also essential to reduce power consumption and improve the profit of service providers by reducing processing time. So, for task scheduling, various "quality of service" (QoS) parameters are considered for reducing execution time and maximize throughput. For this purpose, multi-objective optimization of task scheduling is used, which is a sub domain of "multi-criteria decision making" problem. This involves more than one objective function that can be optimized at the same time.*

Keyword: *Cloud computing, multi-objective, QoS parameter*

Introduction

In the IT industry, Cloud computing [5] is the latest buzzword. With the foundations of grid computing, utility computing, service oriented architecture, virtualization and web 2.0; it is an emerging computing paradigm. With the ownership of just an Internet connection, the user can access all the required software, hardware, platform, applications, infrastructure and storage. A cloud is a type of parallel and distributed system a collection of interconnected and virtualized computer that are dynamically provisioned and presented as one or more unified computing resources based on SLAs established through negotiation between the service providers and consumers. In this information technology oriented growing market of businesses and organizations, cloud computing provides virtual resources that are dynamically scalable. It describes virtualized resources, software, platforms, applications, computations and storage to be scalable and provided to users instantly on payment for only what they use [5].

Multi-objective Optimization

Optimization [5] deals with the problems of seeking solutions over a set of possible choices to optimize certain criteria. They become single objective optimization problems, if there is only one criterion to be taken into consideration, which have been extensively studied for the past 50 years. So, we have multi-objective optimization problems, if there is more than one criterion which must be treated simultaneously. Multiple objective problems arise in the design, modeling and planning of complex real systems in area of industrial production, urban transportation, capital budgeting, forest management, reservoir management, layout and landscaping of new cities, energy distribution, etc. It is easy to see that almost every important real-world decision problems involves multiple and conflicting objectives, which need to be tackled while respecting various constraints, leading to overwhelming problem complexity. The multiple objective problems have been receiving growing interest from researchers with various backgrounds since early 1960. There are a number of scholars who have made significant contributions to the problem. Among them Pareto is perhaps one of the most recognized pioneers in the field.

Scheduling and scheduling criteria

Scheduling is to allocate task to appropriate machine to achieve some objectives or we can say it determines on which machine which task should be executed. In traditional scheduling tasks are directly mapped to resources at one level, whereas, now a days , resources in cloud are scheduled at two level i.e. physical level and VM level which is depicted in Figure1.There are mainly two types of task scheduling in cloud computing: static scheduling and dynamic scheduling. In static task scheduling, information of task is known before execution like execution time whereas in dynamic task scheduling, information of task is not known before execution [11]. In cloud environment to execute a task a user request for a computing resource which is allocate by cloud provider after finding the appropriate resource among existing as shown in Figure 1. Tasks which are submitted for execution by users may have different requirements like execution time, memory space, cost, data traffic, response time, etc. Also, the resources which are involved in cloud computing may be diverse and geographically dispersed. There are different environments in cloud: single cloud environment and multi-cloud environment.

Scheduling process [6] in cloud can be categorized into three stages namely–

a. Resource discovering and filtering – Resource request is made by cloud user and submitted to service provider; service provider searches the suitable resources to locate them.

b. Resource selection –Resource is selected on the basis of task and resource selection parameters.

c. Task submission -Task is submitted to the selected resource.

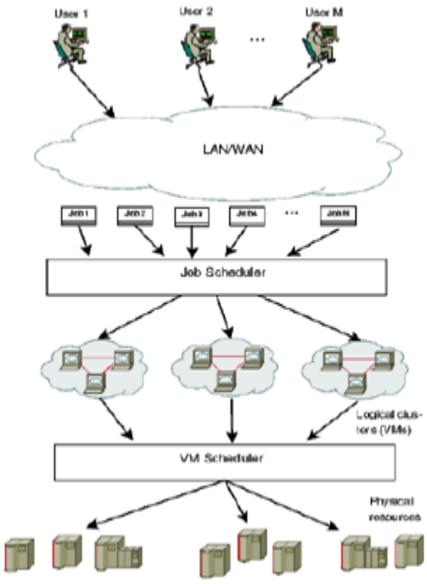


Figure1 :View of cloud

Scheduling criteria

The criteria differ with respect to service provider and user. Service provider wants to gain revenue, maximize resource utilization with minimal efforts, whereas, user wants his job to be executed with minimal cost in minimum time [10]. Figure 2 below shows the Conceptual modeling of the Cloud Computing environment for Task Scheduling.

A. Cloud User Preferred

1) **Makespan**: it tells about the finishing time of last task. The makespan should be minimum, which shows the fastest execution time of a task.

2) **Cost:** it is the sum of the amount paid by the user to provider for using individual resource.

3) **Waiting time**: the time spent by a task in ready queue to get a chance for execution.

4) **Turnaround time (TAT)**: Time taken by a task to complete its execution after its submission i.e., the sum of waiting time and execution time of a task.

5) **Tardiness:** the delay in execution of a task i.e. difference between finishing time and deadline of the task. For an optimal scheduling the tardiness should be zero which shows no delay in execution.

6) **Fairness:** this shows that all tasks are getting equal opportunity of execution.

7) **Response time**: time taken by a system to start responding (first response) after submission of a task.

B. Cloud provider preferred

1) **Resource utilization**: the resources should be fully utilized by keeping them as busy as possible to gain the maximum profit.

2) **Throughput**: his represents the number of task completed in a per time unit.

3) **Predictability:** this represents the consistency in the response times of task .Unpredictable response time may degrade the performance of system.

4) **Priority:** To give preference to a task to finish it as earliest. Priority can be given on the basis of arrival time, execution time, deadline etc. Resources are provided to higher priority task to complete the execution.

5) **Load balancing**: distribution of load among all the computing resources.

6) **Deadline**: the time till which a task should be completed.

7) **Energy efficiency**: Reducing the amount of energy used to provide any solution or service.

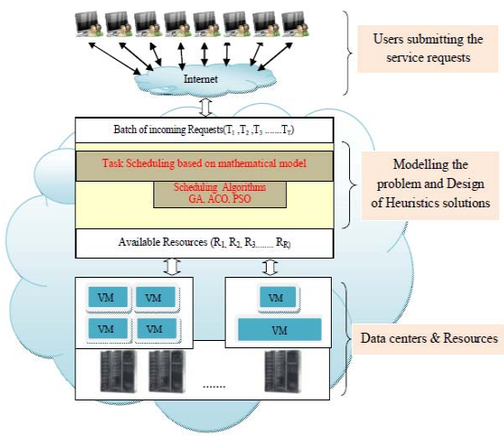


Figure 2 : Conceptual modeling of the Cloud Computing environment for Task Scheduling

Literature Survey

Dynamic Multi-objective task scheduling in Cloud Computing based on Modified particle swarm optimization(2015): A.I.Awad et.al[1] state the efficient allocation of tasks to available virtual machine in user level base on different parameters such as reliability, time, cost and load balancing of virtual machine. Agent used to create dynamic system. The proposed mathematical model multi-objective Load Balancing Mutation particle swarm optimization (MLBMPSO) is used to schedule and allocate tasks to resource which is

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

shown below in Figure 3. MLBMPSO considers two objective functions to minimize round trip time and total cost. Reliability can be achieved in system by getting task failure to allocate and reschedule with available resource based on load of virtual machine.

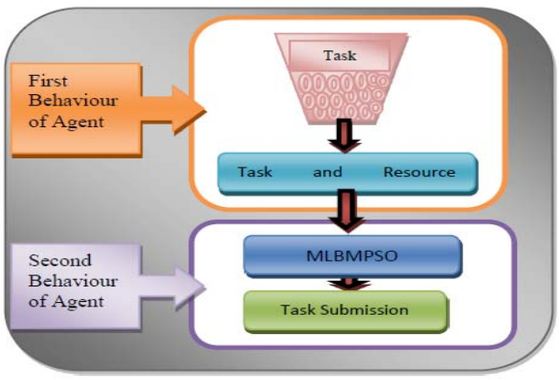


Figure 3: Proposed model by author

Figure 3[1] depicts two behaviors:
First behaviour is responsible for
1. Task Buffer
2. Task and Resource Information
Second behaviour is responsible for
1. MLBMPSO
2. Task Submission

Responsive Multi-objective Load Balancing Transformation Using Particle Swarm Optimization in Cloud Environment (2016): VG.Ravindhren.et.al [9] states that Resource allocation among multiple clients has to be ensured as per SLAs. So, to accomplish the goals and achieve high performance, it is important to design and develop a Responsive multi-objective load balancing Transformation algorithm (RMOLBT) based on abstraction in multi cloud environment. The model is represented below in Figure 4. It is the most challenging to schedule the tasks along with satisfying the user's Quality of Service (QoS) requirements. This paper proposes a wide variety of task scheduling and resource utilization using Particle swarm Optimization (PSO) in cloud environment. The results in this paper demonstrate the suitability of the proposed scheme that will increase throughput, reduce waiting time, reduction in missed process considerably and balances load among the physical machines in a Data centre in multi cloud environment.

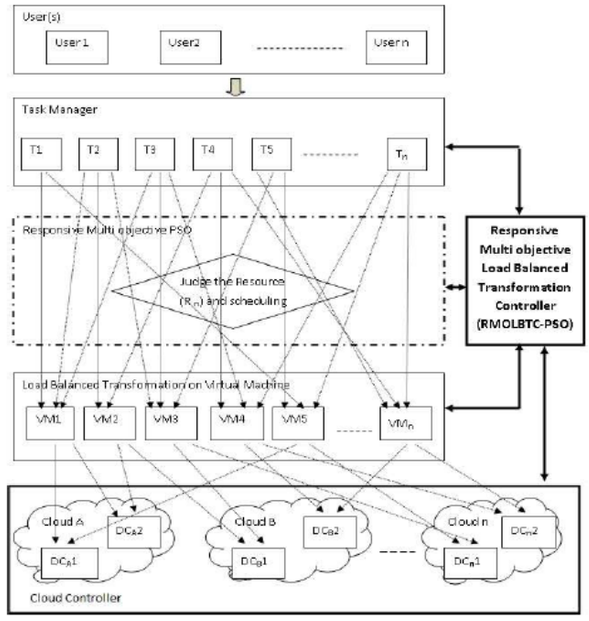


Figure 4: System Model for Responsive Multi objective Load Balance Transformation using PSO

An Efficient Approach for Task Scheduling Based on Multi-Objective Genetic Algorithm in Cloud Computing Environment (2014): Sourabh Budhiraja et.al[5] state that the scheduling of the cloud services to the consumers by service providers influences the cost benefit of this computing paradigm. In such a scenario, tasks should be scheduled efficiently such that the execution cost and time can be reduced. In this paper, the author proposed an efficient approach for task scheduling based on Multi- Objective Genetic Algorithm (MOGA) shown below in Figure 5, which minimizes execution time and execution cost as well. For task scheduling, a Multi-Objective genetic algorithm [5] is implemented and the research is focused on crossover operators, mutation operators, selection operators and the Pareto solutions method.

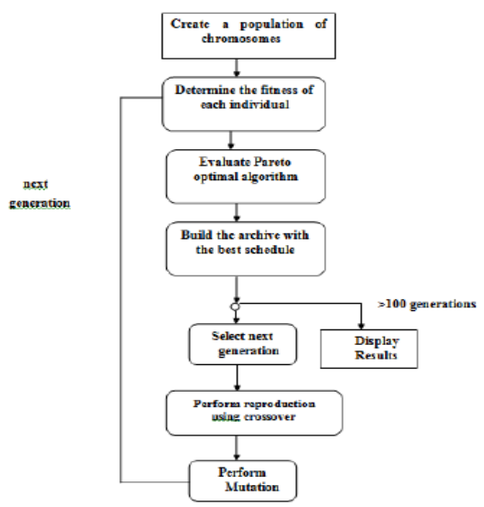


Figure5: Flow Chart-Modified Genetic Algorithm(MGA)[1]

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Multi-Objective Tasks Scheduling Algorithm for Cloud Computing Throughput Optimization(2015): Atul Vikas Lakra et.al[2] state that the problem is to bind set of tasks received by the broker to the received list of VMs, so that execution time of workload is reduced to minimal optimized time. Single objective scheduling algorithms have some problem. For example, in priority task scheduling, high priority tasks always get chance to execute, due to this low priority task have to wait for a long time. Sometimes low priority task gets a chance to execute but, if high priority tasks keep coming then low priority task is pre-empted and CPU is allocated to high priority task and this leads to increase in execution time of a task as well as it reduces the throughput of the system. Similarly, in First Come First Serve (FCFS) and Shortest Job First(SJF) ,task scheduling algorithms face problem in worst case scenario. These algorithms perform very well in the best case but in worst case they degrade the performance to very low level. So an efficient scheduling algorithm is required which can provide optimized performance in both cases. Using a proper scheduling algorithm implementation in broker improves the datacenter's performance without violating service level agreements. The order of task submission and the VMs also influence the execution time of the entire workload.

.

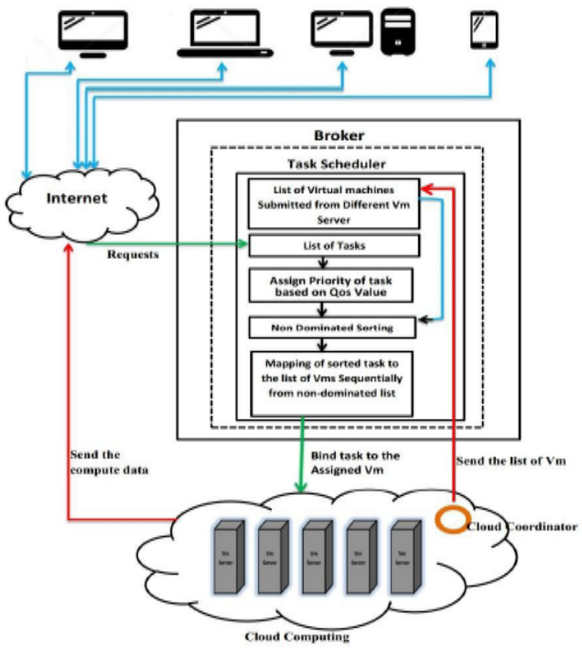


Figure6: Cloud computing architecture

The cloud computing architecture and the proposed work is shown above in Figure 6, which depicts that the Cloud broker is responsible for mediating negotiations between SaaS and cloud provider and such negotiation are driven by QoS requirements. Broker acts on behalf of SaaS for allocation of resources that can meet application's QoS requirements. Also, Figure 7 below represents the algorithm of multi –objective task scheduling [5] proposed by the author.

Algorithm 1 : Multi-objective task scheduling algorithm

1. Submit both VMs list of successfully created VMs in datacenter and task list to Broker.
2. Create a received list of tasks.
3. Create a received list of VMs.
4. Non-dominated sorting (list of task)

```
i ← 0
Create empty non-dominated list
dominated list ←list of task
Initially put task_i in the non-dominated list
for all i ← 1 to size of task's list do
  for all j ← 0 to size of non-dominated list do
    if task_j dominates task_i then
      put task_j into non dominated set
    else
      if task_i dominates task_j then
        put task_i into non dominated set
      end if
    else
      put task_i and task_j into non dominated set
    end if
  end for
end for
```

5. Sort the list of task according to the non-dominated task set.
6. Sort the VM received list in descending order .
7. $j \leftarrow 0$ 7.

```
for all i ← 0 to the size of task's list do
  if j ≥ 0 then
    Bind task_i to the VM_j j++
    if j== number of VMs then
      j=0
    end if
  end if
end for
```

Figure 7: Algorithm for multi-objective task scheduling[2]

Multi-Target Tasks Scheduling Algorithm for Cloud-environment Throughput Optimization(2016): Shubhashree S. R et.al[7] proposed that proposed multi-task scheduling algorithm that enhances the data center execution without damaging SLA. The proposed algorithm is as appeared below in Figure 8 and Figure 9, that utilize non-dominating sorting algorithm for comprehending the multi-objective (task size, QOS value). After a fixed time, interval, the list will be updated dynamically. This algorithm will give the optimized throughput when compared with the existing algorithm.

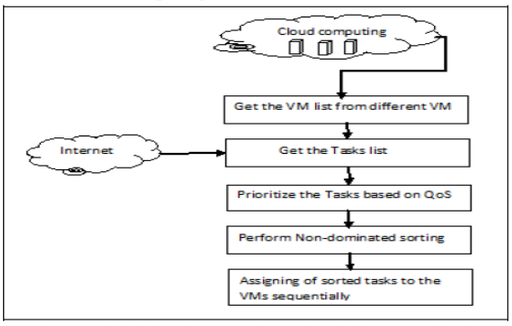


Figure8 :Multiobjective task scheduling and dominance relation

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

1. Submit both VMs list of successfully created VMs in datacenter and t
Broker.
2. Create a received list of tasks.
3. Create a received list of VMs.
4. Non-dominated sorting (list of task)
$i \leftarrow 0$
Create empty non-dominated list
dominated list ←list of task
Initially put $task_i$ in the non-dominated list
for all $i \leftarrow 1$ to size of task's list **do**
  for all $j \leftarrow 0$ to size of non-dominated list **do**
    if $task_j$ dominates $task_i$ **then**
      put $task_j$ into non dominated set
    else
      if $task_i$ dominates $task_j$ **then**
        put $task_i$ into non dominated set
      end if
    else
      put $task_i$ and $task_j$ into non dominated set
    end if
  end for
end for
5. Sort the list of task according to the non-dominated task set.
6. Sort the VM received list in descending order .
7. $j \leftarrow 0$ 7.
for all $i \leftarrow 0$ to the size of task's list **do**
  if $j \geq 0$ **then**
    Bind $task_i$ to the VM_j j++
    if j== number of VMs **then**
      j=0
    end if

Figure 9: Algorithm for multi-target task scheduling algorithm[7]

Multi-Objective Task Scheduling in Cloud Computing Using an Imperialist Competitive Algorithm(2016): Majid Habibi.et.al[3] states that Since the tasks scheduling in the cloud computing environment and distributed systems is an NP-hard problem, in most cases to optimize the scheduling issues, the meta-heuristic methods inspired by nature are used rather than traditional or greedy methods. One of the most powerful meta-heuristic methods of optimization in the complex problems is an Imperialist Competitive Algorithm (ICA). Thus, in this paper, a meta-heuristic method based on ICA is provided to optimize the scheduling issue in the cloud environment. Figure 10 below depicts the pseudo code proposed by the author in this paper.

Input:npop(Population-size),problem-size,ep,α, β, pr
For i=1 **to** npop **do**
  $C_{iposition}$ ← RandomPosition(problem-size)
  If i<=ep then
    EmpiresPopulation ← $C_{iposition}$
  Else
    C_w← GetWorstSolution(EmpiresPopulation)
    If Cost($C_{iposition}$) < Cost($C_{wiposition}$) **then**
      Replace(EmpiresPopulation,C_i,C_w)
    Else
      $C_{iempire}$ ← ssignAnEmpire(EmpiresPopulation)
    End
  End
  Populaton ← C_i
End
EvaluatePopulaton(Population)
EvaluateEmpiresPopulation(EmpiresPopulation,Population)
ImperialisticCompetition(EmpiresPoplution,Population)
EliminiateWeakestEmpire(EmpiresPoplution,Population)

End
EvalutePopulation(Population)
BestSol ← GetBestSolution(Population)
Return BestSol

Figure 10: Pseudo code for the algorithm proposed by the author [3]

Multi-Objective Task Scheduling using K-mean Algorithm in Cloud Computing(2016): Vanita Dandhwani.et.al[8] states that K-mean clustering algorithm is used to create the clusters for tasks. In which for k clusters centroids are calculated based on multi-objectives Task length and Deadline using equation (1) and (2) and Centroid is calculated using equation (3) where minimum distance value is selected as centroid.

Tl= Number of Instructions (MI)--- (1)
Dl= VMmips / Tl -----(2)
Where Tl=Tasksize
Dl=Deadline
VMmips= MIPS of Average VM
dist((x, y), (a, b)) = $\sqrt{(x - a)^2 + (y - b)^2}$----- (3)
Where x= tasksize
y=deadline

Figure 11 below shows the k-means multi objective task scheduling algorithm and resource selection algorithm

A. *K-MEAN ALGORITHM*

Step1: select k points as initial centroid.
Step2: Repeat
Step3: Form k cluster by assigning each point to its closest centroid.
Step4: Recompute the centroid for each cluster.
Step5:Untill centroid do not change.

B: *MULTI-OBJECTIVE TASK SCHEDULING ALGORITHM*

Step1: Get a list of unscheduled task.
Step2: Create a cluster using K-mean algorithm.
Step3: Arrange clusters in descending order (higher the centroid higher the cluster).
Step4: Map clusters to the Best VM using Resource selection algorithm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

C: *RESOURCE SELECTION ALGORITHM*

Step1 : Input: Get Resource list.

Step2: Begin i=0

Step3: While cluster[i] is not empty do

Step4: Select the VM which has máximum capacity using equation (4) [12].

$$Ci=Pro_{ni} * P_{mpi} + VM_{bwi} \quad (4)$$

Where Pro_{ni} is the number of processors in VMi,

P_{mpi} is millions of instructions per second of all processors in VMi

Vm_{bwi} is the communication bandwidth ability of VM_i

Step5: Schedule the cluster and execute it.

Step6: Update status of resorces.

Step7: i=i+1

Step8: End while

Step9: End.

Figure11: Algorithm proposed by the author [11]

Multi objective Task Scheduling in Cloud Environment Using Nested PSO Framework (2015): R K Jena.et.al[4] focuses on task scheduling using a multi-objective nested Particle Swarm Optimization(TSPSO) to optimize energy and processing time.

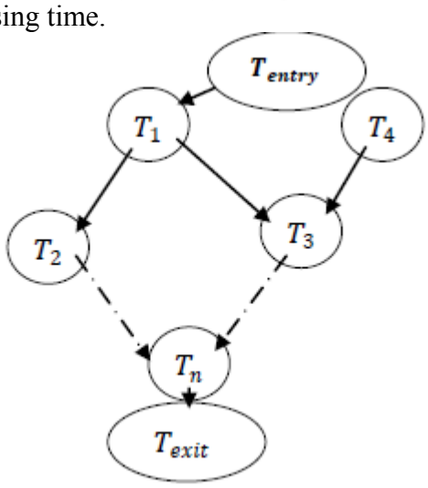


Figure12: Task graph

Figure12 above represents Each vertex V in the DAG is associated with a value <l>, 'l' represents the length of the task in Million Instruction (MI). The problem of this model is how to optimally schedule user jobs to the Processing Elements available in the cloud under different data center. All the PEs is considered homogeneous, unrelated and parallel. Scheduling is considered as non preemptive, which means that the processing of any task can't be interrupted. Figure 13 represents the algorithm proposed by the author of this paper.

Algorithm MOPSO( )
{
Initialize External Archive (Æ) // Æ =∪ *E*
For j = 1 to M (*M* is the size of particle swarm)
Initialize S_j & V_j // Initialization of each particle swarm and its velocity
For k=1 to L // L is the number of iteration
{
For j = 1 to M
E[j] = PSO (S_j)// E [j] is the archive for particle swarm S_j
Update the archive (Æ) of non-dominated solutions

Select leader particle from the archive (Æ)//
Update velocity
Update position
}
Return (Non-dominated solution)
}

Algorithm PSO(S_j)
{ // S_j . represents the set of user tasks of j^{th} particle allocated to different data center (D_t) ,
t = 1,2...P
For t = 1 to P //P is the number of available data center in Cloud environment
{
For i = 1 to N (*N* is the size of particle swarm)
{
Initialize S[i]
Initialize the velocity *V* of each particle V[*i*]
Initialize the Personal Best *pBest* of each particle. *pBest*[*i*] = S[*i*]
Evaluate objectives of each particle: *Evaluate S[i]*
Initialise the Global Best particle (*gBest*) with the best one among the 'N' particles:
gBest = Best particle found in S
} // end of loop 'i'
Add the non-dominated solutions found in S into *EA[t]* // *EA[t]* is the External Archive storing the pareto front for the task assign to data center D_t
Initialize the iteration number (k) = 0
Repeat until k := G // (*G* is the maximum number of iterations)
{
For i = 1 *to N* (swarm size)
{
Randomly select the global best particle for S from the External Archive *EA[t]* and store its position in *gBest*.
Calculate the new velocity V[*i*] according to (7)
Compute the new position of S[i] according to (8)
If (*i* := *G* **PMUT*) *then* // (*PMUT* is the probability of mutation)
Perform mutation on S[i]
Evaluate S[i] *using (2) and (3)*
Update the personal best solution of each particle S[i]
Update the External Archive *EA[t]*
} // end of loop 'i'
}
Retain the best pareto solution in *EA[t]*
} // end of loop 't'
Return (Min { EA[i].Makespan$_{i=1..p}$}, Sum{ EA[i].Energy$_{i=1..p}$})
}

Figure 13: Proposed algorithm by author [4]

Conclusion:

Above Literature summarizes the multi-objective task scheduling algorithm in one form or the other. As we know, single objective functions cannot fulfill all the criteria, e.g., if we consider only priority of the task , rest of the QoS factors, which are very important in scheduling are left like task length , execution time , deadline , cost etc. So, multi-objective task scheduling algorithm is important for enhancing the cloud environment performance. Below Table1 summarized the algorithms of multi-objective task scheduling proposed by the various authors.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Paper name	Authors' name	QoS parameters	Algorithm used	Which parameter is improved
Dynamic Multi-objective task scheduling in Cloud Computing based on Modified particle swarm optimization(2015)	**A.I.Awad, N.A.El-Hefnawy and H.M.Abdel_kader**	reliability, time, cost and load balancing of virtual machine(VM).	multi-objective Load Balancing Mutation particle swarm optimization (MLBMPSO) algorithm	Execution time and makespan is minimized
Responsive Multi-objective Load Balancing Transformation Using Particle Swarm Optimization in Cloud Environment(2016):	VG.Ravindhren and Dr. S. Ravimaran	Job size	PSO algorithm	increase throughput, reduce waiting time, reduction in missed process considerably and balances load among the physical machines in a Data centre
An Efficient Approach for Task Scheduling Based on Multi-Objective Genetic Algorithm in Cloud Computing Environment(2014):	Sourabh Budhiraja, Dr. Dheerendra Singh	Cost and size	Multi-Objective Genetic Algorithm (MOGA)	minimizes execution time and execution cost
Multi-Objective Tasks Scheduling Algorithm for Cloud Computing Throughput Optimization(2015)	Atul Vikas Lakraa, Dharmendra Kumar Yadav	Size, cost	Multi-objective task scheduling algorithm	better performance and improved throughput, reduced cost
Multi-Target Tasks Scheduling Algorithm for Cloud-environment Throughput Optimization(2016):	Shubhashree S. R	Size of task	*Non-Dominated Sorting algorithm*	*to enhance the data-center throughput,* diminishes the execution time.
Multi-Objective Task Scheduling in Cloud Computing Using an Imperialist Competitive Algorithm(2016):	Majid Habibi, Nima Jafari Navimipour	Task size	Imperialist Competitive Algorithm	Execution time
Multi **objective Task Scheduling in Cloud Environment Using Nested PSO Framework**	R K Jena	Size of task	multi-objective nested Particle Swarm Optimization(TSPSO)	optimize energy and processing time.
Multi-Objective Task Scheduling	Vanita Dandhwani, Dr.Vipul Vekariya	Task length and deadline	multi-objective task scheduling	minimize the execution time

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

using K-mean Algorithm in Cloud Computing(2016):			algorithm using k-mean clustering	and makespan

Table1 :Summary of survey

References:

1. A.I.Awad, N.A.El-Hefnawy and H.M.Abdel_kader," Dynamic Multi-objective task scheduling in Cloud Computing based on Modified particle swarm optimization", Advances in Computer Science: an International Journal, Vol. 4, Issue 5, No.17 , September 2015.
2. Atul Vikas Lakraa, Dharmendra Kumar Yadav ," Multi-Objective Tasks Scheduling Algorithm for Cloud ComputingThroughput Optimization", International Conference on Intelligent Computing, Communication & Convergence, Procedia Computer Science 48 (2015) 107 – 113
3. Majid Habibi, Nima Jafari Navimipour," Multi-Objective Task Scheduling in Cloud Computing Using an Imperialist Competitive Algorithm", *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.*
4. R K Jena," Multi objective Task Scheduling in Cloud Environment Using Nested PSO Framework", Procedia Computer Science 57 (2015) 1219 – 1227.
5. Shikha Chaudhary ,Saroj Hiranwal ,C. P. Gupta," Review on Multiobjective Task Scheduling in Cloud Computing using Nature Inspired Algorithms", *International Journal of Emerging Research in Management &Technology, ISSN: 2278-9359 (Volume-6, Issue-8)*
6. Shubhashree S. R.," Multi-Target Tasks Scheduling Algorithm for Cloud-environment Throughput Optimization",International Journal of innovative research and development,Volume 5 Issue 8,July 2016(ISSN:2278-0211)
7. Sourabh Budhiraja, Dr. Dheerendra Singh, "An Efficient Approach for Task Scheduling Based on Multi-Objective Genetic Algorithm inCloud Computing Environment", JCSC VOLUME 5 • NUMBER 2 JULY-SEPT 2014 PP. 110-115 ISSN-0973-7391.
8. Vanita Dandhwani, Dr.Vipul Vekariya," Multi-Objective Task Scheduling using K-mean Algorithm in Cloud Computing", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 11, November 2016.
9. VG.Ravindhren and Dr. S. Ravimaran ," Responsive Multi-objective Load Balancing Transformation Using Particle Swarm Optimization in Cloud Environment", J o u r n a l o f A d v a n c e s i n c h e m i s t r y, V o l u m e 1 2 N u m b e r 1 5, I S S N 2 3 2 1 - 8 0 7 X.
10. Xhafa F, Abraham A., "Computational models and heuristic methods for Grid scheduling problems", *Futur Gener Comput Syst*, Volume 26, Issue 4, pp. 608–621, April, 2010.
11. Y.Chawla, M.Bhonsle, "A Study on Scheduling Methods in Cloud Computing", *International Journal of Emerging Trends & Technology in Computer Science*, Volume 1, Issue 3, pp. 12-17, September –October, 2012.

Novel Evaluation of Protein Convertase Substilisin/Kexin Type 9 (PCSK9) Gene by Motif Detection Technique

Raiha Tallat
CS.Department
(CIIT) Sahiwal
Sahiwal, Pakistan
raiha.tallat@live.com

M.Shoaib
CS. Department
(CIIT) Sahiwal
Sahiwal, Pakistan
mshoaib@ciitsahiwal.edu.pk

Javed Ferzund
CS.Department
(CIIT) Sahiwal
Sahiwal, Pakistan
jferzund@ciitsahiwal.edu.pk

Ahmad Nawaz Zaheer
MS.Department
(GCUF) Sahiwal
Punjab, Pakistan
anzaheer@hotmail.com

Sana Yaseen
CS. Department
(CIIT) Sahiwal
Sahiwal, Pakistan
sanayaseen42@yahoo.com

Umar Draz
CS.Department
(CIIT) Sahiwal
Sahiwal, Pakistan
sheikhumar520@gmail.com

Abstract— **with the emerging of new development in solving the issues related to big data and its important in computer science cannot be undermined. In this research paper we described briefly the motif occurrence and uniqueness for a significant PCSK9 gene responsible for the generation of protein which causes lower cholesterol levels. Motif occurrence is very common in protein sequences and their detection provides a very important role in evaluating the function of proteins. Currently many motif databases are present which help us comparing the specific motifs with currently available motifs, results in the paper are compared with 11 available databases associated with the TOMTOM tool. Results are calculated using three different clusters of PCSK9 protein sequences. 48 different species are the members of the clusters used in the analysis done by the help of de novo algorithm used by the MEME suite. The approach can be considered as the novel example of the renowned problem of motif detection in large graphs using big data analytic techniques.**

***Keywords*-motifs; PCSK9; MEME; transcription; bioinformatics; Genomics**

I. INTRODUCTION

In the era where data has raised to a huge amount, The emerging size of data is opening up the new possibilities for the data scientists in every field whether it's Mechanics, Bioinformatics, Genomics, Media, Business, Computer sciences, Electronics, Health sciences, Telecommunication etc [1]. As it keeps growing it needs to be analyzed so that the hidden information underneath can be fetched and useful decisions can be made on its base. Literature shows that data scientists have developed many algorithms regarding data visualization and analysis to make it more presentable and interpretable for the sake of information retrieval [2]. Data analysis is overlapped with visual analytics because visual analysis has its significant place in the world of Big Data [3]. One of the appropriate techniques for representing data is in the form of graphs. The nodes in a graph can represent the prominent entities depending upon the type of data being processed. According to literature visual graph analysis has gained the attention of researchers in order to process various data formats. One of the paradigms of Big Data is that few algorithms provide efficient results but with limited scalability of data. In case of massive data sets for example protein sequences and DNA sequences, algorithms with more accuracy are required in order to obtain results. Biologists have been working on massive datasets using the techniques of Big Data analytics [4]. In large dynamic graphs often repetitions of patterns are occurring which show that some specific path is being followed repeatedly. This sub graph or repetitive path is known as the "Network Motif" [5]. In case of protein sequences these repetitions are found abundantly thus motif occurrence becomes a factor of similarity. Repetitive patterns in proteins show the function similarity. This similarity holds a special importance because proteins are mutating with the passage of time which causes various changes in different species. These mutations can be compared thus creating connectivity which identifies the similarity between protein structures [6]. Random projections are used for the discovery of motifs with the help of an algorithm named PROJECTION [7]. Other than PROJECTION an algorithm named de novo motif finder is also currently the shining star in network motif detection thus showing the best motifs found in a sequence and their comparison with other motif databases. Among different gene sequences the PCSK9 gene is very important because this gene is responsible for the generation of protein which lowers the level of cholesterol in the blood stream thus resulting into a decreased rate of cardiac

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

diseases [8]. The PCSK9 helps breaking down the low–density lipoproteins receptors which are carriers of cholesterol in the bloodstream.

II. Related Work

Motif detection is performed using a number of different algorithms using DREME, MAST, GYM, DMINDA[2] etc. The DREME algorithm is used for the discovery of motifs based on transcription factor (TF). It allows large data sets to be analyzed and obtains various binding motifs in sequences .Another program used for motif detection is GYM. GYM is known for the Helix-turn-Helix motif detection. Helix-turn-Helix motifs are among the widely studied motif structures as per literature [9], moreover GYM also provides comprehensive information on the protein sequences. Some approaches of artificial intelligence have also been used for the discovery of motifs in gene sequences, one such example is the MAST algorithm, which conveniently uses output from MEME [10] for searching databases such as SWISS-PROT and Genpret. MAST projects some statistical measures that permit a rigorous evaluation of the significance of database searches with individual motifs or groups of motifs [11]. Another facility for the detection of motifs is DMINDA, now known as DMINDA[2], it's an integrated web browser used for the discovery of motifs in the given sequences. The interface of DMINDA[2] also provides the location of a given motif in a sequence [12]. It provides a suite of *cis*-regulatory motif analysis functions on DNA sequences. DMINDA[2] follows four steps for the DNA sequences analysis:

III. Proposed Methodology

In this section the methodology for the sequence analysis is mentioned. Three clusters of sequence s having 48 different species are downloaded from the UniProt database. All the sequences were in FASTA format and unaligned. The sequences were than uploaded to t he MEME for motif discovery with specific parameters. Minimum width for motif detection was 6 and maximum width w as 50. The sequence of the analysis is represented by the given diagram. Starting from the outer circle and moving towards the target. The goal is to find a motif unique in nature. Identification of a unique motif means that motif can be studied for biologists and it has some specific functionality which is not present in the existing databases.

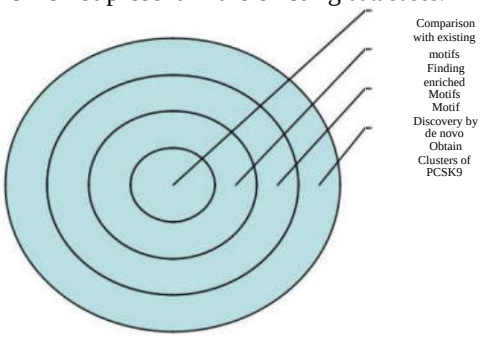


Figure 1: Analysis Methodology

IV. Results And Conclusions

This section represents results obtained from 3 different clusters of PCSK9 gene sequence. MEME suite was used for the discovery, locating enrichment and comparison of motifs.

A. *DeNovo motif detection by MEME*

The figure shown below represents the de novo motif discovery using the MM algorithm and d comparison of motifs using the TOMTOM technique across existing motif databases.

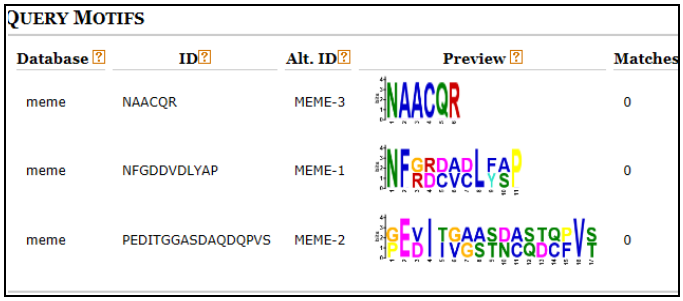


Figure 2: DeNovo motif detection using MM Algorithm

Figure 3 represents the results calculated by the MAST algorithm used in the MEME suite.

Figure 3: Initial and ending point of the identified motif using MAST Algorithm

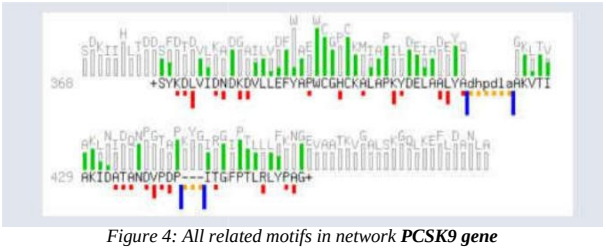


*Figure 4: All related motifs in network **PCSK9 gene***

The figure given below shows the motif with minimum E-value of 1.1e+001 in the PCSK9 protein sequence. The variation in the size of the alphabets representing amino acids in the motif shows conservation.

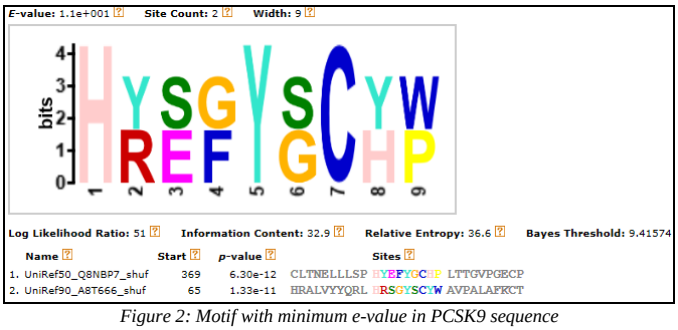


Figure 2: Motif with minimum e-value in PCSK9 sequence

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. The Bayes Threshold:

The bayes threshold of the considered motif is calculated using the Bayesian optimum classifier. Our results showed that the naïve classifier performed efficiently for detecting motifs with the sequence having range of 400 residues. It also performed well with the controlled sequences which were the shuffle version of the original sequences in order to observe motif conservation.

CONCLUSION

The approach can be considered as the novel example of the renowned problem of motif detection in large graphs using big data analytic techniques. After discovering motifs in PCSK9 sequences the next step was determining the uniqueness and enrichment of the motifs. This paper describes the motif occurrence and uniqueness for a significant PCSK9 gene responsible for the generation of protein which causes lower cholesterol levels. Results were concluded with the fact that PCSK9 gene obtained from 48 different species had the smallest motif with length 49 which represents a significant amount of similarity. Further comparison with existing TOMTOM databases showed that only 1 motif was determined unique in nature thus representing the mutation caused by PCSK9. Few motifs had similarity with other motifs but similarity index was limited as compared to other motifs. Detected motifs can be further used for finding the binding pockets of the proteins thus determining its functionality.

REFERENCES

[1] Wernicke, S., &Rasche, F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics*, *22*(9), 1152-1153.

[2] Moses, A. M., Chiang, D. Y., &Eisen, M. B. (2003). Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. In *Biocomputing2004* (pp. 324-335)

[3] Vahdatpour, A., Amini, N., &Sarrafzadeh, M. (2009, July). Toward Unsupervised Activity Discovery Using Multi-Dimensional Motif Detection in Time Series.In *IJCAI* (Vol. 9, pp. 1261-1266).

[4] Xing, E. P., Wu, W., Jordan, M. I., & Karp, R. M. (2004). LOGOS: a modular Bayesian model for de novo motif detection. *Journal of Bioinformatics andComputational Biology*, *2*(01), 127-154..

[5] Wong, E., Baur, B., Quader, S., & Huang, C. H. (2011). Biological network motif detection: principles and practice. *Briefings in bioinformatics*, *13*(2), 202-215.

[6] Dodd, I. B., & Egan, J. B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic acids research*, *18*(17), 5019-5026.

[7] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., &Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, *298*(5594), 824-827.

[8] Dodd, I. B., & Egan, J. B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic acids research*, *18*(17), 5019-5026.

[9] Harrison, S. C., & Aggarwal, A. K. (1990). DNA recognition by proteins with the helix-turn-helix motif. *Annual review of biochemistry*, *59*(1), 933-969.

[10] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, *37*(suppl_2), W202-W208.

[11] Bailey, T. L., Baker, M. E., & Elkan, C. P. (1997). An artificial intelligence approach to motif discovery in protein sequences: application to steroid dehydrogenases. *The Journal of steroid biochemistry and molecular biology*, *62*(1), 29-44.

[12] Ma, Q., Zhang, H., Mao, X., Zhou, C., Liu, B., Chen, X., & Xu, Y. (2014). DMINDA: an integrated web server for DNA motif identification and analyses. *Nucleic acids research*, *42*(W1), W12-W19.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Review on Detection and Counter Measures of Wormhole Attack in Wireless Sensor Network

Rabia Arshad, Saba Zia

***Abstract*- Sensor nodes are organized to form wireless sensor network to be deployed in hostile environments. Sensor nodes communicate each other routing protocols. Information from source node to destination node is sent via intermediate nodes. Security is a major issue in WSN in present days as WSN is vulnerable to attacks that can cause damage to the functionality of the system. In this survey paper an attempt has been made to analyze security threats to WSN at network layer. Network layers is affected by many attacks e.g. Black Hole Attack, Grey Hole Attack, Wormhole Attack, out of which Wormhole attack is the most devastating where attacker agents make a link between two points with low latency. This paper focuses on some researches in detecting and preventing the wormhole attack in network layer.**

Keywords- ***Mobile Adhoc Network, Security, Wireless Sensor Network, Wormhole Attacks***

I. INTRODUCTION

WSN [1-5] is composed of numerous sensor nodes that are capable of monitoring environmental conditions. Sensor nodes are responsible for transmitting the information in the network. Transmitting data in the WSN is a critical task because sensor nodes are restricted devices. Due to this reason, sensor network is susceptible to many attacks. WSNs have some special features that distinguish them from other networks. These characteristics are given as follows [18]:

- Limited resources
- Minimum battery life span
- Self-configuration
- Random changes in topology of network
- Centralized approach of network control

Security is one of the important challenges in designing the WSN. Data is vulnerable to attacks of many kinds therefore security measurements should be taken while designing the WSN. Many security attacks can affect the performance of WSN e.g. Black hole [19], Grey hole [20], Wormhole attack. Wormhole attack is the most dangerous attack to WSN. From this point of view, this paper briefly describes the techniques for detection and prevention of wormhole attack. Rest of the paper is described as follows:

- Section 2 describes about the challenges of WSN.
- In section 3 various attacks on WSN are summarized.
- Section 4 covers the background of Wormhole attack
- Section 5 describes the different types of Wormhole attack
- In section 6, different modes of wormhole attack are discussed
- Section 7 listed some counter measures to wormhole attack

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

II. CHALLENGES OF WSN

According to different application requirements the following design objectives of sensor nodes are considered [16, 17]:

A. Low Cost Node

Sensor nodes are deployed usually in a harsh environment in a large quantity. Also the sensor nodes are not reusable. Therefore, reducing the cost of the sensor nodes is an important step in network design.

B. Low Cost Node

Reducing the size of the sensor node reduces the power consumption as well as also the cost of sensor nodes. Reducing the node size is very useful in the node deployment in hostile environment.

C. Low Energy Consumption

Sensor nodes consume power in performing their function. Power in the sensor nodes is provided by the batteries. In some situations it is impossible to charge their batteries. Therefore reducing the power usage of sensor nodes is a crucial factor. In this way, the network lifetime can be prolonged.

D. Scalability

Routing protocols for the sensor networks must be scalable to different network sizes as the sensor network consist of thousands of sensor nodes.

E. Reliability

Protocols for sensor networks must include error detection and correction techniques. By these techniques, a protocol ensures the reliability of data delivery over some noisy link.

F. Adaptability

In sensor networks, any fault can occur in the network due to which a node may fail. A new node may be added in the network at some later stage. It is also possible that a node may move to some new place in the network (in the mobile network). These situations result in variations in the network topology. Therefore, the network protocols should be adaptive to such changes in the network.

G. Channel Utilization

The network protocols should efficiently use the bandwidth to improve utilization of the channel as the sensor networks have limited bandwidth resources.

H. Fault Tolerance

WSN is mostly deployed in the harsh or hostile environment. Nodes might get failed due to the harsh environmental conditions. Therefore, sensor nodes must have the characteristics of fault tolerant.

I. Security

Information in the network must be secured and prevented from malicious attacks. Thus, effective security techniques must be introduced in the sensor network to avoid these kinds of attacks.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. Attacks on WSNs

WSN are susceptible to a variety of attacks because of multi-hop transmission. As nodes are deployed in hostile environment therefore WSNs have some additional vulnerability to attacks. Table I summarizes the possible attacks on different layers of WSN and their possible solutions.

TABLE I. Attacks on Different Layers on WSN and Their Solutions

Layer	**Attacks**	**Security Solutions**
Physical	Tampering	Tamper Proofing Hiding Encryption
Data Link	Jamming Collision	Error correction method Spread Spectrum Method
Network	Sybil Sinkhole Wormhole	Authentication Authorization
Transport	Packet Injection	Packet Authentication
Application	Aggregation Based Attacks	Cryptographic Approach

IV. Network Layer Attacks and Their Effects

Layered architecture of WSN make it more vulnerable to security attacks. Various attacks and their defensive techniques have been proposed in WSN. Attacks on network layer [6] are given in Table II.

TABLE II. Network Layer Attacks In WSN and Their Effects

Attack	Description	Effects
Wormhole	a. Require 2 or more adversaries b. These adversaries have better resources of communication between them [7].	• Network topology changes. • Packets are destructed. • False information for routing.
Sybil	A malicious node represents different identities and attracts the traffic [8].	• Disruption of WSN • Can be a source for other attacks
Black Hole	A malicious node behaves like destination node and does not forward the packet.	• Throughput is decreased [7] • Disruption of WSN
Sink Hole	More complex than Black Hole [7]	• Attracts all the traffic • Other attacks are also triggered. • Base station position is affected.
Selective Forwarding	A malicious node selectively drops the packet and does not forward. It acts like Black Hole [9].	• Message contents are modified. • Packet dropping • Resources can be exhaustive.
False Routing	Here the attackers creates the loops in the network by routing packets to false sink node [10].	• False messages • Resource exhaustion

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. Background and Significance of Wormhole Attack

WSN is vulnerable to attacks of different types due to the scarcity of resources. Wormhole attack is a severe attack on network layer of WSN where two or more attacking agents are connected by high speed wormhole off-channel link [12]. Wormhole attack has two different mode of attacking i.e. 'Hidden' and 'Exposed' mode. In exposed mode of attack, identity of attacker is attached in the packet header while tunneling and replaying packets [11].

In wormhole attack, any two attackers form a tunnel to transfer data and replays this data in the network. This tunnel is referred to as wormhole. Wormhole attack effects the WSN tremendously. Routing techniques may be disrupted when routing messages are tunneled. Figure. 1 represent a scenario of wormhole attack. Packets received at node A are replayed via node B and vice versa.

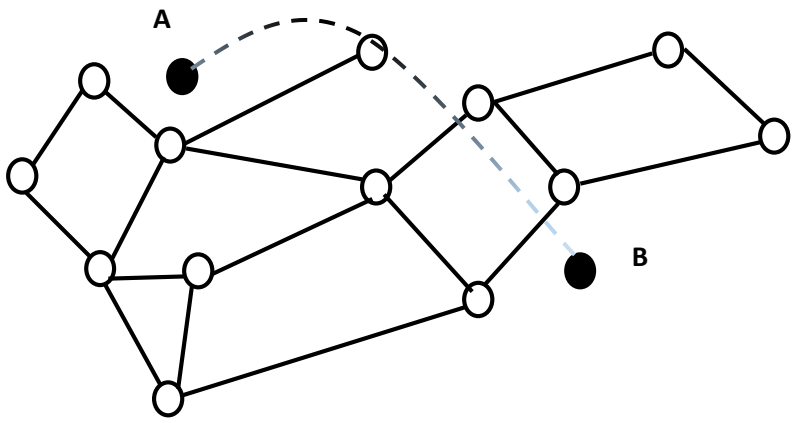


Figure 1. Wormhole Attack

VI. Classification of Wormhole Attack

Wormhole attack is very difficult to detect. It is one of the Denial of Service attacks. Wormhole attack might be launched by any number of nodes. It is required to categorize the wormhole attacks for detection and prevention of wormhole attacks. Wormholes are divided into three categories [13] i.e. *Open, Closed and Half Open*. Table III describes which nodes are visible or invisible in three types of classifications. Figure 2, 3 and 4 describe the three types of wormhole attack respectively where S= source, D= destination, M= Malicious Node. Classification of wormhole attack is based on following:

i. Attackers are invisible/ visible
ii. Data forwarding mechanism of wormhole nodes
iii. Ability to hide and show the identities of nodes

TABLE III. Nodes Description in Different wormhole Attacks

	Open	**Half Open**	**Closed**
Source Node	Visible	Visible	Visible
Destination Node	Visible	Visible	Visible

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Malicious Node1	Visible	Visible	Invisible
Malicious Node2	Visible	Invisible	Invisible

A. Open Wormhole

Source node, destination node and malicious nodes are visible in open wormhole. Nodes X and Y are kept hidden on traversing path. In this mode, packet header also contains the attackers. All sensor nodes in the network are aware of the presence of malicious nodes and would represent the malicious nodes as direct neighbors.

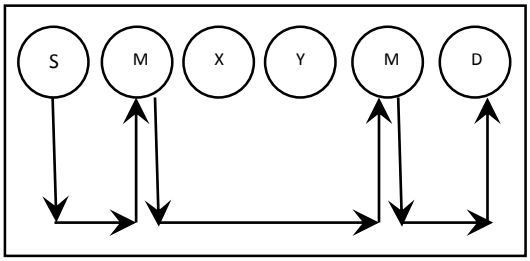


Figure 2. Open Wormhole

B. Closed Wormhole

In this mode, source node and destination node behaves one hop away from each other and this leads to creation of fake neighbors.

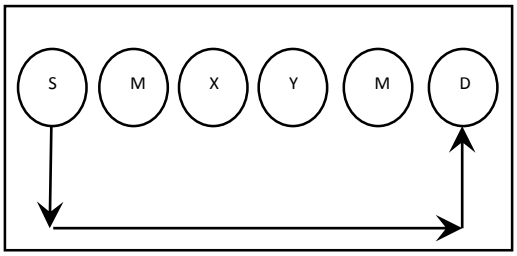


Figure 3. Closed Wormhole

C. Half Open Wormhole

The first end of malicious node is visible near source node and second end is hidden in this mode of wormhole attack. Contents of the data packet are not modified by attackers. Packets are rebroadcast in this type of attack because the attackers tunnel the packets from one malicious end to other.

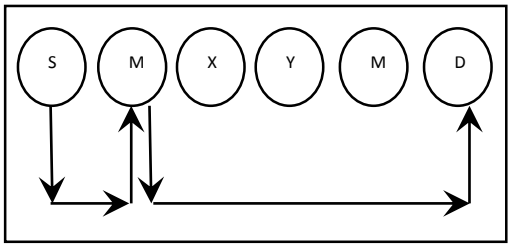


Figure 4. Half Open Wormhole

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VII. Modes of Wormhole Attacks

Wormhole is categorized into the following types based on how many nodes are involved in establishment of wormhole.

A. Wormhole Using Protocol Distortion

This type of wormhole does not affect the network functioning too much so it is a harmless type. It is also known as "Rushing Attack". Only a single node is a malicious node that distorts the routing protocol.

B. Wormhole using Packet Relay

In this type of attack, malicious nodes replay data in between two nodes at far distance from each other. This would lead to creation of fake nodes. It is also "Replay Based Attack".

C. Wormhole using Out-of-Band Channel

In this type of attack, there is only one malicious node capable of high transmission. It attracts the data packets to traverse the route that passes from the malicious node.

D. Wormhole using Packet Encapsulation

In this type of attack, there are many nodes presents between any two malicious nodes. Data is encapsulated in between malicious nodes.

Table IV summarizes different modes of wormhole attack and number of adversary nodes is given [13]:

TABLE IV. Summary of wormhole Attack Modes

Mode	Minimum Adversary Nodes
Protocol Distortion	1
Packet Relay	1
Out of Band Channel	2
Packet Encapsulation	2

VIII. Counter Measures of Wormhole Attack

In this section, some important wormhole detection methods are discussed. Table V summarizes a description of wormhole detection methods.

A. Statistical Analysis Method

Song et al. proposed a mechanism for detection of wormhole based on statistical analysis of multipath routing. This method is useful for multipath and on-demand routing protocols in intrusion detection systems [14].

B. Graph Theory Method

Graph theory method was proposed by Lazos and Poovendran [11] for detection of wormholes. In this method, nodes that are used have information about location and thus are called location-aware guard nodes (LAGNs). LAGNs use a key called "local key" between neighbors that are one hop apart. The message that is encrypted with local is not possible to decrypt. Therefore hashed messages are used in formation of local key to detect wormholes. Node will detect a number of inconsistencies in the message if there is a wormhole otherwise it will unable to response.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

C. Hop Counting Method

Hidden as well as exposed attacks can be detected using this method. In this method, DELPHI [14] (Delay per Hop Indicator) protocol is used. Length of each route and time delay for each route is calculated for identification of wormhole. Therefore, the route having wormhole faces a greater time delay as compared to other routes.

D. Visualization Based Method

In this method [11], each node computes the distance to its neighboring nodes by using received signal strength. According to these calculations, networks topology is calculated by the base station. Network topology is more or less flat in case there are no wormholes. In wormholes are present then in visualization it can be seen there is a string at different ends of network that pulls it. In this method, each node has to send its list of neighbors to the base station.

E. Hardware Based Method

A method based on directional antennas [15] was proposed by Hu and Evans. It was based on assumption that in absence of any wormhole, if node A sends data in particular direction than it was received at its neighbor from opposite direction. It is mandatory that every node must have its directional antennas.

F. Trust Based Method

Jain and Jain proposed another important method [14] for identification and isolation of malicious nodes that create wormhole in the network. In this method, trust values are calculated according to the sincerity of the nodes in the neighbors for execution of routing protocol. The trust value is used to effect the routing decisions which guide the nodes not to communicate through wormholes. Packet dropping is reduced to 14% by using the trust based methods. Using trust based mechanisms, throughput is also increased up to 8-9%.

TABLE V. SUMMARY OF METHODS FOR WORMHOLE DETECTION

Methods	Description
Statistical Analysis Method	Efficient for on demand protocols, easy integration and effective for multipath routing
Graph Theory Method	Nodes are equipped with GPS receivers
Hop Counting Method	Low overhead, high efficiency and fast performance
Visualization Based Method	Mobility is not considered in this method. It is best for dense networks
Hardware Based Method	Not applicable to the networks other than having directional antennas. Very efficient for the networks having antennas
Trust Based Method	Locate dependable routes in the network in an effective way

IX. CONCLUSION

In this paper, attacks to WSN are studied briefly. Wormhole attack is one of the important attacks on WSN. Background of wormhole attack, its cause's and methods of prevention i.e. graph theory method, trust based method, and hop-counting methods etc. are discussed in this paper. Graph theory method requires GPS receivers it is location based method. Hardware method is the efficient method for large networks having antennas. Hop-counting method have high efficiency over other methods.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] Neha Rathi, Jyoti Saraswat and Partha Pratim Bhattacharya, "A REVIEW ON ROUTING PROTOCOLS FOR APPLICATION IN WIRELESS SENSOR NETWORKS". *International Journal of Distributed and Parallel Systems,* Vol. 3, No. 5, September 2012

[2] Kemal Akkaya, Mohamed Younis, "A Survey on Routing Protocols for Wireless Sensor Networks". *Ad Hoc Networks* (2005)

[3] Abdulaleem Ali Almazroi, MA Ngadi, "A Review on Wireless Sensor Networks Routing Protocol: Challenges in Multipath Techniques" *Journal of Theoretical and Applied Information Technology*, Vol. 59, No. 2, January 2014

[4] Rajashree, V. Biradar, V. C. Patil, Dr. S. R. Sawant, Dr. R. R. Mudholkar, "Classification and Comparison of Routing Protocols in Wireless Sensor Networks". *Special Issue on Ubiquitous Computing Security Systems.*

[5] Laiali Almazaydeh, Eman Abdelfattah, Manal Al- Bzoor, and Amer Al- Rahayfeh, "Performance Evaluation of Routing Protocols in Wireless Sensor Networks". *International Journal of Computer Science and Information Technology*, Volume 2, Number 2, April 2010

[6] Nityananda Sarma , Sangram Panigrahi, Prabhudutta Mohanty and Siddhartha Sankar Satapathy, "Security Issues in Wireless Sensor Network Data Gathering Protocols: A Survey". Journal of *Theoretical and Applied Information Technology*, 2005

[7] Kia Xiang, Shyaam Sundhar Rajamadam,Srinivasan, Manny Rivera,Jiang Li, Xiuzhen Cheng, "Attacks and Countermeasures in Sensor Networks: A Survey", *Springer*, 2005

[8] Sushma, Deepak Nandal, Vikas Nandal, "Security Threats in Wireless Sensor Networks". *International Journal of Computer Science & Management Studies*, Vol. 11, Issue 01, May 2011

[9] Syed Ashiqur Rahman, Md. Safiqul Islam, "Anomaly Intrusion Detection System in Wireless Sensor Networks: Security Threats and Existing Approaches". *International Journal of Advanced Science and Technology*, Vol. 36, November 2011

[10] Ali Modir Khazeni, Norafida Ithnin, Mohammad Javad Abbasi, "Secure Hierarchal Routing Protocols in Wireless Sensor Networks: Security Survey Analysis", *International Journal of Computer Communications and Networks*, Volume 2, Issue 1, February 2012

[11] Majid Meghdadi, Suat Ozdemir and Inan Guler , "A Survey of Wormhole Based Attacks and Their Countermeasures in Wireless Sensor Networks", *IETE TECHNICAL REVIEW*, Vol. 28, ISSUE 2, April 2011

[12] Khin Sandar Win, "Analysis of Detecting Wormhole Attack in Wireless Networks". *World Academy of Science, Engineering and Technology*, 2008

[13] Marianne Azer, Sherif El-Kassas, Magdy El-Soudani, "A Full Image of the Wormhole Attacks Towards Introducing Complex Wormhole Attacks in Wireless Ad Hoc Networks", *International Journal of Computer Science and Information Security*, Vol. 1, May 2009

[14] Preeti Nagrath, Bhawna Gupta, "Wormhole Attacks in Wireless Adhoc Networks and their Counter Measurements: A survey". *IEEE,* 2011

[15] Zhibin Zhao, Bo Wei, Xiaomei Dong, Lan Yao, Fuxiang Gao, "Detecting Wormhole Attacks in Wireless Sensor Networks with Statistical Analysis". *IEEE International Conference on Information Engineering*, 2010

[16] Jamal Al-Karaki and Ahmed E. Kamal, "Routing Techniques in Wireless Sensor Networks: A Survey". *IEEE Communications Magazine*, Vol. 11, No. 6, December 2004

[17] Kemal Akkaya and Mohamed Younis. "A Survey on Routing Protocols for Wireless Sensor Networks", *Ad hoc Networks*, Vol. 3, No. 3, May 2005

[18] Shio Kumar Singh, M P Singh, and D K Singh, "Routing Protocols in Wireless Sensor Networks – A Survey". *International Journal of Computer Science & Engineering Survey,*, Vol. 1, No.2, November 2010

[19] Arshdeep kaur, Mandeep kaur, "A Survey Black Hole Attack in Manet". *International Journal of Science, Engineering and Technology Research*, Volume 4, Issue 5, May 2015

[20] Mitali Khandelwal1, Sachin Upadhyay, "Detecting and Preventing Black hole and Grey Hole Attacks for Trust Management in Wireless Sensor Networks: A Survey". *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, Issue 5, May 2016

Performance Analysis of spot diffusion technique for indoor optical communication system employing LDPC

Sabira Khanam Shorna
Department of Computer Science & Engineering,
Primeasia University,
Dhaka, Bangladesh
Email: sabira001840@gmail.com

Koushik Chandra Howlader
Dept. of CSTE,
Noakhali Science and Technology University,
Noakhali, Bangladesh
Email: koushiksoul@gmail.com

Arup Mazumder
Department of Electronics and Telecommunication Engineering,
Southeast University,
Dhaka, Bangladesh
Email: arup.mazumder27@gmail.com

***Abstract*— Indoor optical wireless communication systems offer an attractive substitution for realizing next generation wireless local area network. In the context of indoor optical-wireless communication (OWC) system the spot-diffusing technique provides improved performance compared to conventional diffuse system. In this work, the performance of an OW spot-diffusing communication system using Neuro-Fuzzy (NF) adaptive multi-beam transmitter configuration has been proposed. This research work focus on the performance of the indoor optical wireless systems. Where BER of the existing modulation technique evaluated based on the Spot Diffusion Adaptive ANFIS algorithm and compared the existing result and analysis. A Linear error correcting code low-density parity-check (LDPC) scheme is introduced in the algorithm to observe the change in the performance of the system. It has been analyzed that involving LDPC in Spot Diffusion Adaptive ANFIS algorithm provides approximately 37.5% improvement of BER result against receiver position, where BER reduce chronologically. From the bit error rate analysis of these schemes it has seen that the system model with LDPC performs better than other existing techniques.**

***Keyword:* - Indoor optical wireless communication (OWC), Neuro-Fuzzy (NF) adaptive multi-beam transmitter, Spot Diffusion Adaptive ANFIS algorithm, Linear error correcting code low-density parity-check (LDPC).**

I. INTRODUCTION

This work is focusing on the performance analysis for the BER of indoor optical wireless communication system using LDPC. Some of advantages of OWC are low cost, base-band circuit design, High data rates (Gaps), less multi access interference, no need to pay for spectrum license etc. and limitations are it can't pass through wall, sensitive to blocking; limited Transmit Power etc. These research works have their own features, possibilities and limitation which are described later in the paper. However wireless systems have some working and performance limitation. Some key points are discussed below:
Optical fiber communication transmits information through optical fibers is largely replaced by radio transmitter systems for long-haul optical data transmission. Not only telephony but also Internet traffic, long high-speed local area networks (LANs), cable TV (CATV) has been used such systems. OWC is defined as the use of optical frequencies to carry the electrical signals. Unguided visible, infrared (IR), or ultraviolet (UV) light to carry a signal.

For indoor communication though infrared provides significant advantages as a medium but it also has some drawbacks. Several aspects impair the performance of indoor IR transmission systems. Because of such hinders design and implementation is not so easy using infrared. So, the optical wireless communication is used to overcome such indoor environment.

Following table 1.1 gives a comparison between optical wireless and radio frequency systems.

Criteria	Optical wireless systems	Radio systems
Bandwidth	Unregulated large	Limited
Passes Through wall	No	Yes
Cost	Low	High
Speed	High	Low
Fading	Free from fading	Multipath fading.
Security	Security and freedom from spectrum regulation and licensing.	Low security.

Table 1.1: Comparission Between Optical Wireless and Radio System in Indoor Wireless Communication.

This can be used for different ranges:
- Short range (cm – m): Chip-to-Chip Interface,
- Medium range (m – 10 m): Wireless Optical LAN,
- Long range (km): Free-Space Optical Communications.

However, In Optical wireless LAN where it can be used to illuminate the room while serving as a medium for data transfer and also for transferring data at high speed for long distance. For free space communication, it also offers high speed long distance data transfer using satellite.

1. Eye safety consideration puts limit on the amount of optical power that should be emanated by the transmitter, thus limiting the coverage of an optical wireless system.
2. In indoor optical wireless systems, the leading source of noise is ambient light, which is typically a combination of fluorescent light, sunlight, and incandescent light. Ambient light provokes shot noise due to the random nature of the photo-detection process.
3. A multipath phenomenon occurs when the transmitted signal follows different paths on its way to the receiver due to its reflection by walls, ceilings and other objects. Channel dispersion associated with multipath propagation is another major issue in indoor optical wireless systems. Multipath phenomena can cause inter-symbol-interference (ISI).

During implementing the performance of optical wireless system in indoor context, some problems were noted. Especially evaluation of BER is major factor for such systems.

another technique named LDPC have used for long-haul data transmission using the principal of forward-error

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

correction in the optical wireless communication. Another work where using Neuro fuzzy technique analysis the BER according to the mobility of the transmitter and receiver. But here BER have not investigated for the respected distance between transmitter and receiver.

Thus, investigating the BER for the allowing distance between transmitter and receiver in indoor context of optical wireless communication is the major issue of this proposed work, where reason of using the LDPC technique is to improve the BER performance then previous work and study.

- Previous work of indoor optical wireless communication has analyzed mainly the concern of BER. All the information that are related to this proposed research are investigate and simulate in Matlab platform.
- If the simulation could have been done in real environment, more appropriate result might have been measured.
- Here considered the upward light illuminating measurement. The measurement in a room for the side wall reflection are not listed.
- Five parameters are considered by the proposed research, they are SNIR (Signal –to-Noise-Interference), Delay Spread, BER with and without LDPC. There are more parameters that could have been compromised.

The objectives of the study are:

- to review on existing indoor optical wireless communication system improvement of BER
- to propose a method like LDPC in this area that look up and analysis the BER for the system.
- to compare the performance of proposed method with existing research

All of this aim would cover the enough pretty analysis of BER for high-speed indoor OW applications.

In chapter 2, a number of different proposed works are discussed based on research papers. Proposed system model has been discussed in chapter 3. Numerical analysis and comparison with other methods have been established in chapter 4.

II. LITERATURE REVIEW

In this chapter, there are brief knowledge about the OWC and BER. Many studies were introduced using this method for different wireless communication from time to time. Here many research are discussed and show comparison among them.

Today's world research heeds their interest mostly in the context of wireless communication. The lofty maintenance and configuring cost makes the wireless system financial and flexible alternative to wired system. People are aware of a proliferation of many researches and developments that has been done in this area specifically in Optical wireless communication. Further generation of wireless communication systems are already relay on Optical Wireless System technologies. Conventionally OWC technology depends on high power solid-state lasers or diode lasers for medium to long-range applications. Latterly we have seen the remarkable advances in semiconductor sources such as light emitting diodes (LEDs) in visible light and ultraviolet wavelengths, multi-array light sources and detectors, tracking and steering. With the benefits of low power and cost for short/medium range wireless communication applications such advances provide huge latent. The optical wireless Communication could provide a cost effective, flexible, secure and ultra-high-speed solution to the materializing challenges facing the system and service providers over RF (Radio frequency).

Optical wireless communications have becoming an effective alternative medium to optical fiber, and radio frequency (RF) communications and it optimistically removes gradually all of the possible hinders that are challenging for previous communication technologies, because of its high bandwidth, low cost, ease of implementation, license free spectrum freedom from interference and many more. For some wireless applications such as 3-D face-to-face communication and super Hi-Vision/Ultra High Definition TV data (more than 4Gbit/s) [1] due to the explosive bandwidth which can be envisioned through optical wireless communication. Data throughput and transmission link based on optical wireless these are the great concern for number of applications [2] [3].

Most of the time the transmitted data it is not fully secure or error free. Some statistical fluctuations are related to noise influences (e.g. laser noise, amplifier noise, shot noise, or excess noise of a receiver) cause a small fraction of the transmitted bits to be defective. Typically, the bit error rate (i.e., the fraction of incorrectly transmitted bits) is strongly dependent on the transmitted power, and the latter must be high enough to keep the bit error rate below a certain acceptable level (e.g. 10^{-12} for Earth-based telecommunication systems, or 10^{-6} for satellite control). Nearly all of the remaining bit errors can be detected using some kind of checksums and corrected.

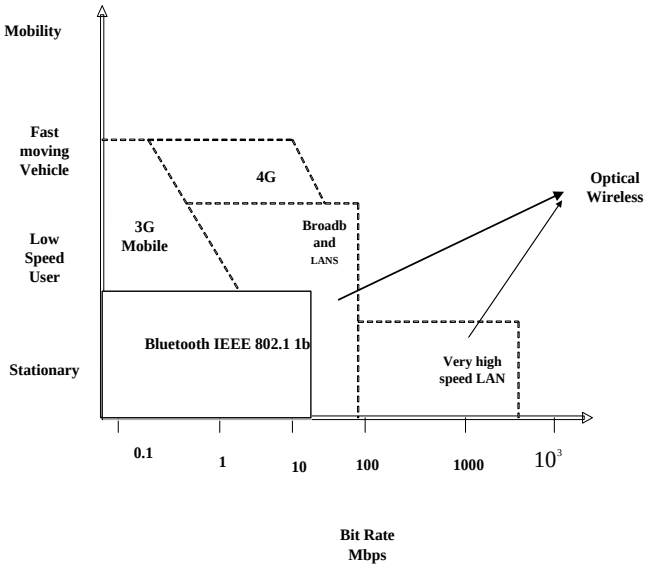


Fig 2.1: Simple Scenario of BER in Optical Wireless communication

The error correction scheme can use some level of redundancy in the transmitted data or involve retransmitting corrupted data packets. Additional detrimental influences such as fiber losses or various types of dispersion in a longer link, or background light in free-space transmission, can often be compensated by somewhat increasing the transmitted power. BER is the total number of erroneous bits compared to the total number of transmitted bits.

The increase of optical power required to maintain a given bit error rate is called a *power penalty*, or more specifically e.g. a *dispersion penalty* if dispersion is the considered factor. The above figure 2.1 presents the scenario of BER related in OWC.

For reducing the probability of loss of information LOW density parity check code (LDPC) is an effective error correcting code that used in noisy communication channel. By employing LDPC this probability can be reduced to as small as desired thus the data transmission can be as close to Shannon's limit.

Optical wireless communication is a dynamic area of research and development. The basic scheme of OWC systems communications and applications of are discussed in [4]. With the main focus on the indoor deployment scenarios, here reviewing and summarizing recent advancements in OW communication [5].

For future high performance and cost-effectiveness of indoor optical wireless systems and possible configurations for modulation, and multi-access techniques are presented in [6].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

An overview of both operate term in the short- (personal and indoor systems) and the long-range (outdoor and hybrid) regimes and discussed within the context of point-to-point ultra-high-speed data transfer. The discussion on outdoor systems focuses on the impact of atmospheric effects on the optical channel and associated mitigation techniques that extend the realizable link lengths and transfer rates. [7]. The authors have discussed about the generic propagation models for the design of wireless optical communication in [5]. Bit error rate of optical wireless communication had addressed by employing a detector array in the presence of turbulence [8]. Pervasive work and study have done to improve performance of the OWC systems through BER by occupying several methods. These articles focused and discussed on the BER of wireless optical communication presenting different methods and approaches [9] [10] [11].

The artificial neural network (ANN) has some attractive properties like adaptability, parallel processing, and universal approximation. A well train ANN can perform the task equalization efficiently. In this paper authors have presented the ANN for equalization in indoor environment in an optical channel [12]. Another work has done where authors described their research about the improvement of the optical wireless system using Neuro-Fuzzy based Spot-Diffusing Techniques which provides better performance compared to other generic spot-beam diffusion method [13]. However, in optical wireless communications LDPC method is effective way to improve the BER (bit error rate).

Many works have done focused on the LDPC for optical wireless communication. This paper investigated the performance of regular and irregular, random like and structured generalized low-density parity-check (GLDPC) codes for long-haul transmission [14]. The authors evaluated the low-density parity check (LDPC) coded schemes with unequal transmission power allocation (UTPA) in optical wireless channel [15]. This work proposed a method of establishing a secure and reliable communication link using optical wireless communication (OWC) [16]. Here implemented a modified Low-Density Parity-Check (LDPC) codec algorithm in ultraviolet (UV) communication system [17]. [18] Here authors analysis the BER of Optical Wireless Communication (OWC) System employing Neuro-Fuzzy (NF) based spot diffusion system. But this system has not yet implemented for the performance of BER using LDPC. However, in recent years the researches of LDPC for indoor optical wireless communication are increasing day by day.

Now comparing the proposed work where introduce the LDPC methods focusing on the improvement of bit error rate analysis for more secure and reliable data transmission in indoor optical wireless communication area.

III. SYSTEM MODEL & PROPOSED METHOD

In this chapter, an established Spot Beam algorithm using LDPC has been proposed. Here, SNIR, Delay Spread, BER has been considered. This research is worked on the existing Spot Beam Selection algorithm using LDPC methods for high speed data transfer and reduce the probability of loss of information.

An Indoor Optical Wireless System: RF (Radio Frequency) and Infrared are two major transmission technologies to gain indoor optical wireless communication [19]. Its performance depends on the propagation and type of system used. The basic systems types can be categorized into diffuse or line of sight (LOS) systems [6], [7]. To obtain high data rates such as Gbit/s can be achieved in LOS systems, [8], [11], but due to its directionality the system is vulnerable to blockage/shadowing. Whereas several paths from source to receiver exist in diffuse OW system, which makes the system robust to blockage/shadowing. However, the path losses are high and multi-paths create inter-symbol interference (ISI) which limits the achievable data rate [8] [11].

Transmission Techniques: Basic optical wireless system consists of a transmitter (using LEDs or LDs). Propagation medium like free space and the receiver (using APDs or PIN diodes). Some transmission techniques are described below:

Directed beam infrared (DBIR) radiation: From the transmitter to the receiver the optical beam travels directly without any reflection. The optical wireless link using this technique is established between two fixed data terminals with highly directional transmitter and receiver at both ends of the link. Lack of mobility and susceptibility to blocking and shadowing by personnel and machines is the main drawback of this technique.

Diffuse infrared (DFIR) radiation: The transmitters send optical signals in a wide angle to the ceiling and after one or several reflections the signals arrive at the receivers in DFIR. For transmission, the system does not require any line of sight alignment which is one of the most desirable configurations from a user point of view. Though, systems using this technique have a higher path loss than their DBIR counterparts, requiring higher transmitter power levels and receivers with larger light collection area. Multipath dispersion is another challenging problem in this technique. Radiate optical power over a wide solid angle. Thus, provide mobility to the receiver.

Quasi-diffuse infrared (QDIR) radiation: In QDIR, there is a base station (BS) with a relatively broad coverage made of passive or active reflector which usually accumulated on the ceiling. By always maintaining the line of sight BS transmits (receives) the signal power to (from) the remote terminals (RTs) thus, the RTs cannot be fully mobile. From any position in the room to enable communication between itself and the BS the RT's transceiver must be aimed to the BS, or its FOV must be wide enough. In another appearance of QDIR technique, the transmitter may send the optical signal to a designated area on the ceiling and the receiver is supposed to face that area. In general, this architecture provides a concession between the DFIR and DBIR option. In QDIR system inherits aspects of both point-to-point and diffuse links. Gradually deviating beam sources which illuminate a grid of spots on the ceiling.

Channel propagation: A number of transmission techniques are possible for indoor optical wireless systems; these techniques may be classified according to the degree of directionality of transmitter and receiver [19]. For indoor OW links, the two most common configurations are LOS and non-LOS propagation systems. In the field of indoor mobile application LOS links rely upon a direct path between the transmitter and receiver, regardless of their beam angles, while non-LOS links generally rely upon light reflections from walls, ceilings and other diffuse reflecting surfaces a non-directed non-LOS link scenario which is often referred as a diffuse link is the most desirable approach. In the environments where shadowing exists diffuse systems also play a very significant role. One of the main reasons is the facts that diffuse propagation systems do not require transmitter-receiver alignment or line-of-sight and instead make use of reflections from walls, ceiling, and other reflectors. However, diffuse systems are subject to multipath dispersion which results in signal spread and inter symbol interference. Through the use of diversity and/or equalization the effects of multipath propagation can be reduced.

System Model: Here we consider an empty room with floor dimensions of 8× 4 m^2 ceiling height of 3m. Where the reflection coefficient of the ceiling is considered to be 0.8. The ceiling has eight spot lights. In the Figure x_0 and x are the position of the imaging receiver and v is the velocity where α is elevation angle δ, is the azimuth angle, d = 8, w = 4 and h = 3. Neuro-Fuzzy (NF) adaptive multibeam

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

transmitter is located at the center of the room whereas an imaging receiver is placed at

x_0 = (1; 1; 0:5).

With adapted beam angle (α, δ) and power that is reflected by the imaging receiver while transmitter generates multi spot beam matrix on the ceiling. Through the low rate diffuse channel, the transmitter learns receiver position, mobility. At low data rate, the beam maintains the fixed power.

Analysis of SNIR and BER: The ambient light affects signal-to-noise-plus interference (SNIR) at the receiver in indoor optical-wireless communication. Many researchers have considered intensity modulation with direct detection (IM/DD) as most feasible approximation. The received signal, symbolize by $y(t)$ can be expressed as:

$$y(t) = \sum Rx(t) * h(t, \alpha, \delta) + \sum n(t, \alpha. \delta) + \sum I(t, \alpha. \delta) \dots\dots (1)$$

Where R is the receiver responsivity, $x(t)$ is the instantaneous optical transmitted power, $h(t, \alpha, \delta)$ is the impulse response of the OW channel, $n(t, \alpha. \delta)$ is the ambient light noise, $I(t, \alpha. \delta)$ is the instantaneous interference power. The SNIR, denoted by γ, of the received signal can be calculated by [9].

$$\gamma = \frac{R^2 (P_{s1} - P_{s0}) \times h^2}{(\sigma_{s1} - \sigma_{s0})^2} \dots\dots\dots\dots (2)$$

Where P_{s1} and P_{s0} are the optical power associated with the binary 1 and binary 0 respectively, σ_{s1} σ_{s0} are the shot noise variation component with P_{s1} and P_{s0} respectively.

Bit Error Rate: For the non-encoded system with binary phase-shift-keying(BPSK), the BER expression can be given by:

$$\psi_{bpsk}(\gamma) = \frac{1}{\pi} \int_0^{\pi/2} \exp\left(-\frac{b_{bpsk}\gamma}{sin^2\varphi}\right) \dots\dots\dots\dots (3)$$

Where $b_{bpsk} = sin^2(\pi/2)$. By (2) and (3) we can write

$$\psi_{bpsk}(\gamma) = \frac{1}{\pi} \int_0^{\pi/2} \exp\left(-\frac{sin^2(\pi/2)\left(\frac{R^2(P_{s1} - P_{s0}) \times h^2}{(\sigma_{s1} - \sigma_{s0})^2}\right)}{sin^2\varphi}\right) \dots (4)$$

Adaptive Power Allocation: The achievable data transmission rate, denoted by *b*, of the OWC system is given by $b = \frac{1}{M} \sum_{i=1}^{M} log_2 \left(1 + \frac{R^2 \times ((P_{s1i} - P_{s0i}) h_i^2}{(\sigma_{s1i} - \sigma_{s0i})^2}\right) \dots (5)$

The optimization problem and constraint of the power allocation can be written as

max $b \dots\dots\dots (6)$

s.t $\sum_{j=1}^{J} P_j \leq \bar{P} \dots\dots\dots (7)$

Where $\bar{P}$ is the average power. To analyze the above optimization problem, we can use the Lagrange multiplier method and the Lagrangian function is defined as $L = b + \mu_j (\sum_{j=1}^{J} P_j \leq \bar{P}) \dots\dots\dots (8)$

where μ_j is the Lagrange multiplier. After solving the Eqn. (8), we can write

$$P_j = \left[\frac{(P + \sum_{j=1}^{J} \frac{1}{h_i})}{C} - \frac{1}{h_i}\right] \dots\dots\dots (9)$$

$$= max\left[\lambda(C) - \frac{1}{h_i}, 0\right] \dots\dots\dots (10)$$

Delay Spread: The Doppler spread of an impulse is expressed as rms value by,

$$D = \sqrt{\frac{\sum (t_i - \mu)^2 P_r^2}{P_r^2}} \dots\dots\dots (11)$$

Where $\mu = \frac{t_i P_r^2}{P_r^2}$ and t_i is the delay time and Pr is the received power. Now considering the Doppler shift and the Neuro fuzzy *ANFIS* model and using the algorithm for Spot Beam selection of this paper.

Proposed System Method: Since the aim of this proposed work is to transmit correct information at longer distance with lower transmit power, LDPC coding scheme is used with coded modulation to achieve significant coding gain without bandwidth expansion.

System model can be explained by using LDPC where, information typically in the form of digital data, is input to electronic circuitry that modulates the transmitting light source (LEDs/LDs). The source output passes through an optical system into the free space (propagation medium). The received signal also comes through the optical system and passes along the optical signal. The optical signal traverses through the indoor channel where it gets affected by the multipath propagation and back ground noise. The noise affected optical signal is detected by the photo detector.

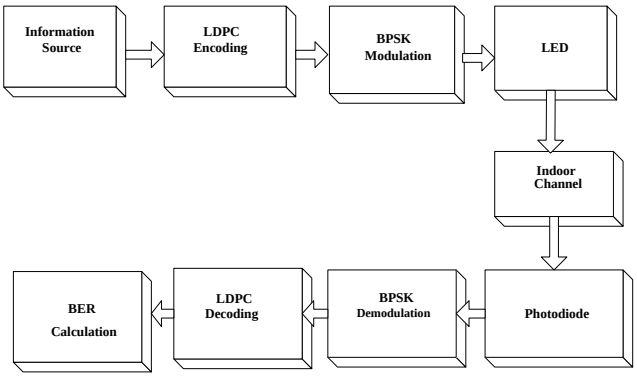


Fig. 3.1: Block Diagram for the LED based Indoor Communication System Model using LDPC coding with BPSK Modulation

Then BPSK demodulation takes place followed by LDPC decoding. The BER analysis is done to estimate the performance of the system.

Error Correction Coding Scheme – LDPC: LDPC codes are based on a parity-check matrix where each transmission message contains (*m*, *n*) binary block with 2^n binary *m* tuples (*m* > *n*). Here *m* is block length, *n* is number of data bits and (*m* − *n*) is number of checked bits. Code rate, ε = $\frac{n}{m}$.

Encoding: The Parity-check matrix, *H*, for (7, 4) LDPC block code is constructed using hamming codes as constituent codes so that the position where error occurs can be detected for correction.

$$H = \begin{vmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{vmatrix} \dots\dots\dots \textbf{(12)}$$

With the associated data bits of parity locations, Generator matrix, G, is generated and given below:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$G = \begin{vmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{vmatrix} \quad \dots\dots\dots (13)$$

Given a message *u*, the codeword *c* will be the product of *G* and *u*.

$c = G.u \dots \dots \dots (14)$

If the received codeword is *v*, the syndrome vector, denoted by *z*, will be

$z = H.v \dots \dots \dots (15)$

When $z = 0$, codeword is error free. Otherwise the value of z is the position of the flipped bit.

Decoding: By sending messages through the connected edges between variable nodes $b_i \varepsilon B$ and check nodes $f_i \varepsilon F$ of factor graph where *B* and *F* are the set of all message nodes and check nodes respectively decoding is accomplished. Here assume that m_{fibi} is the message sent by node b_i to m_{fjbi} vice-versa. Let, the probability that there is a 1at each variable node b_i is a_i where *i* = 1, 2, 3, ⋯ *n*. Now, the probability that there is an even number of 1's on variable nodes is,

$$P_{even} = P_r[b_1 \oplus \dots \oplus b_n]$$

$$= \frac{1}{2}[1 + \prod_{i=1}^{n}(1 - 2a_{bi})] \dots \dots \dots (16)$$

Now, $m_{f_j b_t}(0)$ can be expressed as,

$$m_{f_j b_t}(0) = \frac{1}{2}\left[1 + \prod_{b'_t \varepsilon B \setminus b_t}^{n} (1 - 2a_{b'_t f_t})(1)\right] \dots \dots (17)$$

Therefore, $m_{f_j b_t}(1) = 1 - m_{f_j b_t}(0) \dots \dots \dots (18)$

The variable nodes update their response messages to check nodes using equation (18).

$$m_{f_t b_t}(0) = k_{b_t f_j}\left[(1 - P_i) \prod_{f'_j \varepsilon F \setminus f_j = 1}^{n} \left(m_{f'_j b_t}(0)\right)\right] \dots \dots (19)$$

And

$$m_{f_t b_t}(1) = k_{b_t f_j}\left[P_i \prod_{f'_j \varepsilon F \setminus f_j = 1}^{n} \left(m_{f'_j b_t}(1)\right)\right] \dots \dots (20)$$

Here $k_{b_t f_j}$ is the constant to ensure

$$m_{f_j b_t}(0) + m_{f_j b_t}(1) = 1$$

$$P_i = P_r[b_i = 1|v] \dots \dots \dots (21)$$

At this point, b_i nodes update their decision, denoted by ζb_i ,by calculating the probabilities $Mb_i(0)$ and $Mb_i(1)$ for 0 and 1 respectively.

$Mb_i(0) = kb_i(1 - P_i)\prod_{f_j \varepsilon F} m_{f_j b_t}(0)$ and $Mb_i(1) = kb_i(P_i)\prod_{f_j \varepsilon F} m_{f_j b_t}(1)$.

Therefore, decision, denoted by ζb_i is given below:

$$\zeta b_i = \{1 \; if \; Mbi\,(1) > Mbi\,(0) \; else.$$

The algorithm will terminate if ζb_i satisfies parity-check equation, otherwise predetermined number of iteration will terminate it. Now Algorithm for correct data transmission is summarized as follows: A spot beam scans the ceiling, SNIR and delay spread, $\Delta\sigma$ for each beam have been calculated by the image receiver using Equs (2) and (9). Based on the required minimum SNIR, i.e., min, and maximum delay spread, i.e., max, transmitter selects the spot-beam matrix (H) by NF controller. The transmitter allocates the power for each selected beam adaptively using Eqn (7). Based on Doppler shift, the transmitter adapts the beam angles α and δ. Multi-spot optical transmitter further reduces the by scheduling. Finally, Multi-spot optical transmitter transmits the spot beam matrix to receiver via ceiling. When transmitter achieves the receiver current position then it transmits data u sing LDPC; after getting the data on the other hand decode the data using decoding LDPC code. Go to Step 1 if transmitter gets receiver's position update.

Algorithm: The following algorithm will find the spot beam with an equal power allocation over 40×20 beam hologram or matrix, H.

Notation:

γ =SNIR, σ = delay spread, α = azimuth angle and $\delta =$ elevation angle, T=transmitter, R=Receiver.

P=position of receiver.

Initialization;

Pro1: begin

for each beam i=1 to n do

Calculates γ *and* σ;

Set γ_{min} and σ_{max} ;

if $\gamma =< \gamma_{min}$ *and* $\sigma >= \sigma_{max}$ then

Calculates α and δ;

end

Calculate *P* using ANFIS Controller;

T sends encoded data *c* using LDPC to R located at position *P*;

R decoded data c' and receives it;

end

end

if P of R changes then

Call Pro1

End

IV. NUMERICAL ANALYSIS

The proposed method is worked on the implanted Neuro-Fuzzy based multi beam system (NFMS) diversity receiver configuration. It has been also compared with other spot-beam diffusion method. The research has been done by employing LDPC code into the above investigated method. Thus, Performance for the indoor optical wireless communication improves and shows possible low error rate in the purpose of data transmission.

Result and Discussion: For the analysis of the implemented algorithm, the simulation parameters those are considered and implanted: length, width and height of the room are 8m, 4m and 3 m; the reflection coefficient of the ceiling is ρ= 0:8. One transmitter which is located at (2; 4; 1) location; there is also one receiver; the area acceptance semi-angle of each photo-diode are$2cm^2$ and 65^0 respectively. The number of pixel at the receiver is 200 (with area of 0:01cm2) Pedestrians move typically at the speed of 1 m/s. If the SNIR is computed after 10 μs; there are 8 spot lamp in the room which are located at (1; 1; 1), (1; 3; 1),(1; 5; 1),(1; 7; 1),(3; 1; 1), (3; 3; 1), (3; 5; 1), and (3; 7; 1); and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the wavelength of the light is 850nm. The results of the Adaptive Spot Beam Selection applying ANFIS model have been implanted by using this parameter.

In this section, the numerical results are analyzed using the mathematical equations derived in the previous Chapter and are simulated using the codes written in MATLAB. The performance analysis is based on a scenario for a typical indoor environment in the presence of additive white Gaussian noise. The performance of the system depends on the receiver sensitivity. The 80 ms adaptation time will give overhead of 8%. Adaptation time depends on environment. Receiver computes the SNIR and delay spread and sends this information via a low rate channel to the transmitter. ANFIS consider two inputs. Iterative training of the ANFIS has been done to achieve the desired output. After a predefined simulation time to obtain the simulation result and use them to train.

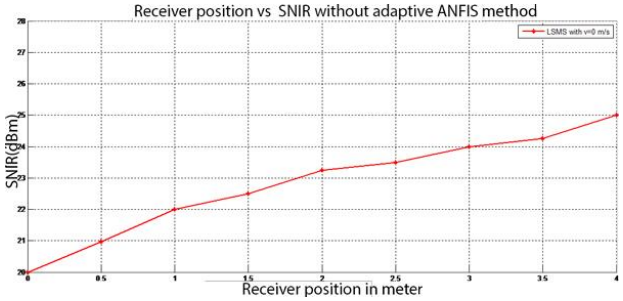


Figure 4.1: Effect of receiver position on SNIR distribution for without Adaptive Neuro Fuzzy System

Another approach is placed here for the Adaptive method is shown in the figure below:

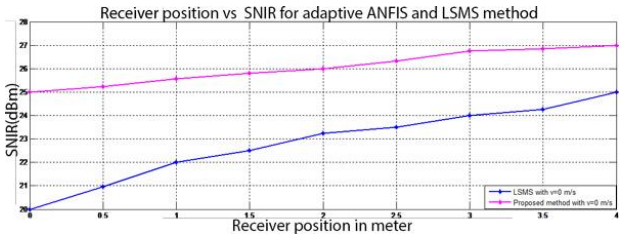


Figure: 4.2: Effect of receiver position on SNIR distribution for Adaptive ANFIS Method and LSMS

From this figure, it seems that here indicates a comparison between proposed model of Adaptive Neuro Fuzzy Method and line strip multi-spot diffuse system (LSMS). The BER calculations were performed for the receiver moving towards the transmitter (i.e., the value of x is increasing) while neglecting the movement along y axis. Significant BER improvement is observed as the NFC moves the spot beam, selects the best positioned spot only, and allocate the power adaptively based on the channel condition of the selected slots. It is also found that the BER performances have been degraded as the receiver is moving. The BER increases as the velocity of the receiver increases.

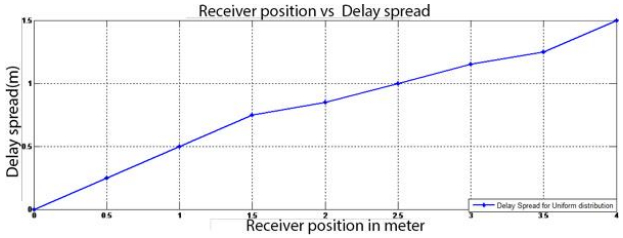


Fig 4.3: Effect of receiver position in Delay Spread

This figure shows the BER comparison for proposed model with LSMS and conventional diffuse system. Performance evaluation revels that BER has been improved, if we change the spot beam angle adaptively.

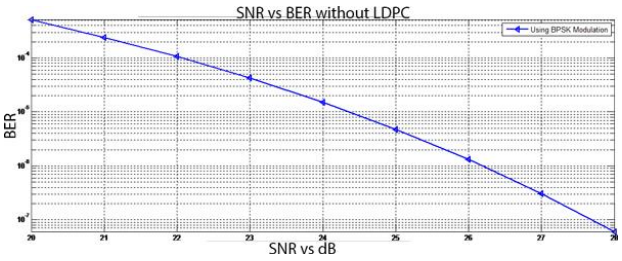


Fig 4.4: SNIR and BER Using BPSK Modulation

This figure described the SNIR and BER performance Using BPSK Modulation technique.

This shows the BER changes in response to the SNIR.

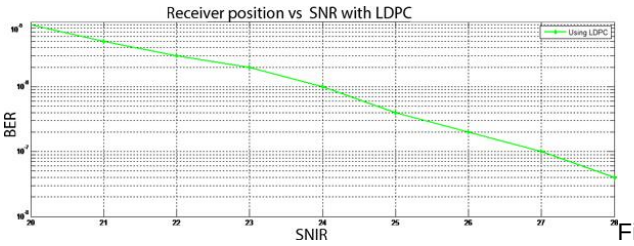


Figure 4.5 represents effects of SNIR and BER using LDPC. Where it indicates the BER reduce in response to the SNIR.

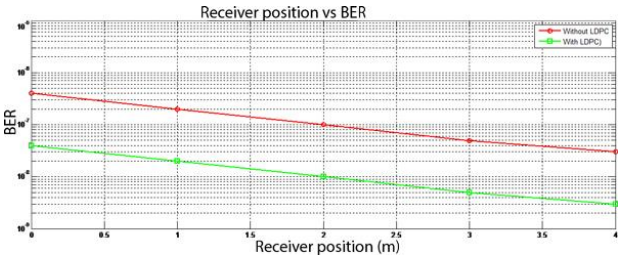


Fig 4.6: BER for Corresponding Receiver position

From the figure 4.6 it has been established that while using LDPC coding scheme then it reflects the better performance. This graph describes the comparison between the performance for the non-coded system and coded (LDPC) system. In which BER consistently increases. Where comparatively LDPC Gives enough pretty result that the proposed method expected.

So, from the above simulation and analysis the study for the desired result has been investigated. Thus, it describes that the Adaptive Spot Diffusion Beam Selection algorithm if coded with LDPC then generates better BER performance.

VII. CONCLUSION AND RECOMMENDATIONS

It is possible to conclude that, in spite of the advances achieved so far, there is still a lot of work to be done to exploit completely the advantages and the potential offered by the optical medium. For indoor wireless system applications, the use of optical communication offers an important alternative or the growing area of mobile computers and communication. Thus, techniques to improve the operation and speeds of infrared wireless systems with room environments have still to found, while trying to decrease the cost of the systems as much as possible. Researcher and manufacturers are also trying to find ways to improve the data bit rates and the range offered by current systems. This work evaluates the performance of a LDPC coded for indoor optical wireless communication system to increase the low bit error rate which can be used for secure and reliable response communication where quality of service is the main concern. Simulation results indicated that BER performance has been improved by incorporating LDPC coding. This work can be extended by considering the mobility and BER of multiple transmitters and receivers using this coding scheme.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] Nagatsuma T., T. Tkakda, H. J. Song, K. Ajito, N. Kukutsu, and Y. Kado, “Millimeter- and THz-wave photonics towards 100-Gbit/s wireless transmission," Proc. of *IEEE Photonics Society's 23th Annual Meeting,* 385-386, Denver, Colorado, USA, 2010.

[2] F.R. Gfeller and U. H. Bapst, “Wireless in-house data communicationvia diffuse infrared radiation”, *Proc. IEEE,* vol. 67, no. 11, pp.1474-1486. Nov. 1979.

[3] T. Komine, M. Nakagawa, “A Study of Shadowing on Indoor Visible-Light Wireless Communication Utilizing Plural White LED Lightings. *Int. Sympo. On Wireless Commun*”, pp. 36-40, 2004.

[4] J. Sánchez-López, A. A. M, F. J. Mendieta and Iván Nieto Hipólito (2011). “Trends of the Optical Wireless Communications, Advanced Trends in Wireless Communications’, Dr. Mutamed Khatib (Ed.), ISBN: 978-953-307-183-1.

[5] Z. Ghassemlooy and A. R. Hayes, “Indoor Optical Wireless Networks”, *IEEE Commun. Mag*., vol. 36, no. 12, pp. 88-94, December 2011.

[6] C. Sing, J. John, Y.N.Singh and K.K.Tripathi, “A Review on Indoor Optical Wireless Systems”, *IEEE Commun. Mag*., vol. 33, no. 15, pp. 79-84, December 2012.

[7] Z. Zhang, L. Dolecek, B. Nikolić, V. Anantharam and M. J. Wainwright “ Design of LDPC Decoders for Low Error Rate Performance”, *IEEE Commun. Mag*., vol. 32, no. 13, pp.57-69, June 2014.

[8] O. Bouchet, M. Bertrand, P. Besnard, “Optical Wireless Communication: LOS/WLOS/DIF propagation model and QOFI software”, *IEEE Commun. Mag.,* vol. 35, no. 17, pp.69-74, December 2010.

[9] Dang A, “A closed-form solution of the bit-error rate for optical wireless communication systems over atmospheric turbulence channels”, 2011 Feb 14; 19(4):3494-502.doi: 10.1364/OE.19.003494.

[10] Aharonovich M, Arnon S, “Performance improvement of optical wireless communication through fog with a decision feedback equalizer”, Opt Soc Am A Opt Image Sci Vis. 2005 Aug; 22 (8):1646-54.

[11] Z. Ghassemlooy, C. K. See and J. M. Holding and C. Lu, “Bit-error-rate analysis for hybrid PIM–CDMA optical wireless communication systems”, 16 AUG 2001 DOI: 10.1002/mop.1351.

[12] A. K. Jaiswal, Anil Kumar, Santosh Tripathi, A. K. C. Shiats, “To Study the Effect of BER and Q-factor in Intersatellite Optical Wireless Communication System”, *IOSR Journal of Electronics and Communication Engineering* (IOSR-JECE) ISSN: 2278-2834, ISBN: 2278-8735. Volume 3, Issue 4 (Sep-Oct. 2012), PP 19-21.

[13] S. Rajbhandari , Z. Ghassemlooy and M. Angelova , “Performance of OOK with ANN Equalization in Indoor Optical Wireless Communication System*”, IEEE Commun. Mag*., vol. 32, no. 14, pp.55-65, March 2010.

[14] I. B. Djordjevic, O. M. and B. Vasic, “Generalized Low-Density Parity-Check Codes for Optical Communication Systems”, *IEEE Commun. Mag*., vol. 37, no. 15, pp.85-95, April 2013.

[15] Sumikawa K. and Habuchi H.,“Optical wireless LDPC-coded PPM with unequal transmission power allocation scheme Oct. 27 2013-Nov. 1 2013 Page(s):44 - 47Print ISBN:978-1-4799-6028-6 INSPEC Accession Number:14431132 Conference Location :Tokyo DOI:10.1109/IWSDA.2013.6849058 Publisher:IEEE.

[16] M. Saadi, A. Bajpai, Y. Zhao, P. Sangwongngam and L. Wuttisittikulkij, “Design and Implementation of Secure andReliable Communication using Optical Wireless Communication”, Frequenz 2014; 68(11–12): 501–509 DOI 10.1515/freq-2014-0027Received February 24, 2014.

[17] M. Wu, D. Han, X. Zhang, F. Zhang, M. Zhang, and G. Yue, “Experimental research and comparison of LDPC and RS channel coding in ultraviolet communication systems” ,published 28 Feb 2014 (C) 2014 OSA 10 March 2014 ,Vol. 22, No. 5 , OI:10.1364/OE.22.005422 ,*OPTICS EXPRESS* 5423.

[18] S. A. Mamun, M. S. Kaiser, M. R. Ahmedz and Md. I. Islam, “BER Analysis of Optical Wireless Communication System Employing Neuro-Fuzzy based Spot-Diffusing Techniques”. “*International Conference on Electrical Information and Communication Technology (EICT),* 2013 ”, ISBN: 978-1-4799-2297-0, DOI: 10.1109/EICT.2014.6777842, Khulna, IEEE.

[19] J. Fernandes, P. A. Watson and J. Neves, “Wireless LANs : Physical properties of Infrared Systems vs Mmw Systems*”, IEEE Commun. Magzine*, August 1994.

Risk Management Process Analysis for Information & Communication Technology (ICT) Systems: Risk Mitigation perspective

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem, Okonkwo, Obikwelu Raphael

[1]Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
Nwaguchikeziekenneth@hotmail.com

[2]Computer Science Department, Michael Okpara University of Agriculture Umudike Umuahia, Abia State, Nigeria.
Saintbeloved@yahoo.com

[3]Computer Science Department, Nnamdi Azikiwe University Awka Anambra State, Nigeria.
Oobi2971@yahoo.com

Abstract— ***This Research Paper, explored Risk management from Risk mitigation perspective. Risk Mitigation Perspective thoroughly examined mitigation processes, Strategies for mitigating both Negative and Positive Risks, discussed essence of Cost-benefit analysis for each recommended control as core mitigation process in order to select cost-effective and best-fit control for the Organization and stakeholders. It further emphasises importance of integration and updating of risk registers in the process and consistent communication/consultation among stakeholders including the Original Equipment Manufacturers (OEMs).***

Keywords—Risk, Risk Management, Risk Mitigation, cost-benefit analysis, Vulnerability, Opportunity

I. INTRODUCTION

In the face of global ever-dynamic threats and attacks, every Organization is adopting measures to reduce negative risks and utilize positive risks. This ensures that her vision and mission is protected, guarded and fully enhanced. This critical as Organizations make ICT, hub for better support and sustenance of her business. As Organizations automated their processes using data and communication devices, Risk Management plays a very critical role in protecting the organizations information assets, and therefore its mission. Risk Management is every stakeholder`s duty and not only for the technical IT team. Therefore, it should be treated as fundamentally as an essential role of the Management. An effective and efficient risk management process is an important component of a successful ICT security so as to ensure data confidentiality, integrity and high availability. According to Project Management Institute (PMI) (2015), Risk is an uncertain event or condition that, if it occurs, has a positive or negative effect on one or more project objectives such as scope, schedule, cost, and quality. A risk may have one or more causes and if it occurs, it may have one or more impacts. **Risk management** is the process of identifying risk, assessing risk, and taking steps to reduce risk to an acceptable level, if possible eradicate it completely. Risk should be both net negative effect of exercise of vulnerability (weakness) according to stoneburner et al (2002) and net positive effect of harness of addendum opportunity (Strength) in the devices in order to fully maximize their utilization and functionality

II. RISK MITIGATION PROCESS

This step involves process of weighing and developing options and actions to enhance opportunities and reduce threats to the devices. Risk mitigation, as the second process of risk management, involves prioritizing, evaluating, and implementing the appropriate risk-reducing or risk-enhancing controls recommended from the risk assessment process. Because the elimination of all risk is usually impractical or close to impossible, it is the responsibility of principal stakeholders to use the least-cost approach and implement the most appropriate controls to decrease mission risk to an acceptable level, with minimal adverse impact on the organization’s resources and mission. Risk mitigation options, an

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

approach for control implementation, control categories, the cost-benefit analysis used to justify the implementation of the recommended controls, and residual risk. Of course, updating of appropriate risk registers.

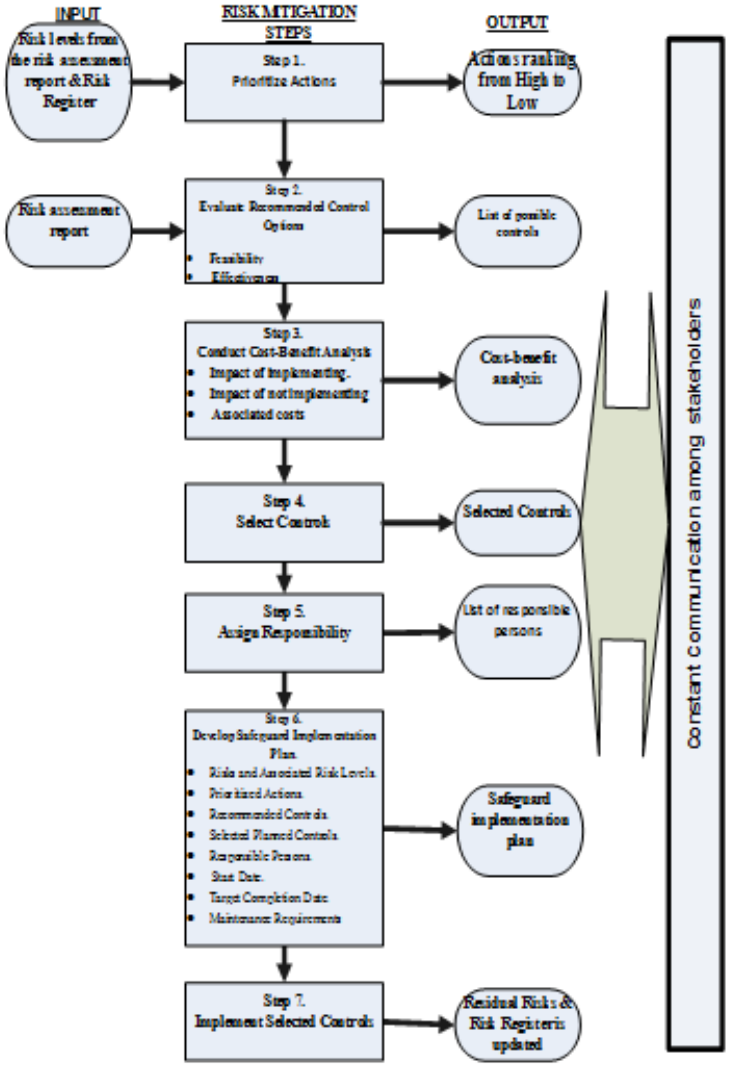


Figure 1: Risk Mitigation Process (*adapted from stoneburner et al*)

Step 1: Prioritize Actions

Based on the risk levels presented in the risk assessment report and content of risk register, the recommended implementation actions are prioritized. In allocating resources, top priority should be given risk items with unacceptably high risk rankings (e.g., risk assigned a Very High or High risk level). These vulnerability/threat pairs will require immediate corrective action to protect an organization's interest and mission.

Step 2: Evaluate Recommended Control Options

The controls recommended in the risk assessment process may not be the most appropriate and feasible options for a specific organization and IT system. During this step, the feasibility (e.g., compatibility, user acceptance) and effectiveness (e.g., degree of protection and level of risk mitigation) of the recommended control options are analyzed. The objective is to select the most appropriate control option for minimizing risk.

Step 3: Conduct Cost-Benefit Analysis

According to Wethli, K. (2014),Cost-Benefit analysis provides one means of identifying the cases in which specific interventions to manage risk do appear to be cost-effective.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

To aid management and principal stakeholders in decision-making and to identify cost-effective controls, a cost-benefit analysis is conducted. The objectives and method of conducting the cost-benefit analysis will be discussed in detail later in this work.

Step 4: Select Control

On the basis of the results of the cost-benefit analysis, management determines the most cost-effective control(s) for reducing risk to the organization's mission. The controls selected should combine technical, operational, and management control elements to ensure adequate security for the IT system and the organization.

Step 5: Assign Responsibility

Appropriate stakeholder (in-house personnel or external contracting staff) who have the appropriate expertise and skill-sets to implement the selected control are identified, and responsibility is assigned by email alert. However, basic user awareness campaign must be carried out company-wide to educate users on what to do and how to do it, including sharing the in-house escalation matrix. All these are to ensure total proactivity in mitigating risks.

Step 6: Develop a Safeguard Implementation Plan

During this step, a safeguard implementation plan (or action plan) is developed. The plan should, at a minimum, contain the following information:

a) Risks (vulnerability/threat pairs) and associated risk levels (output from risk assessment report)
b) Recommended controls (output from risk assessment report)
c) Prioritized actions (with priority given to items with Very High and High risk levels)
d) Selected planned controls (determined on the basis of feasibility, effectiveness, benefits to the organization, and cost)
e) Required resources for implementing the selected planned controls
f) Lists of responsible teams and staff
g) Start date for implementation
h) Target completion date for implementation
i) Maintenance requirements.

The safeguard implementation plan prioritizes the implementation actions and projects the start and target completion dates. This plan will aid and expedite the risk mitigation process.

Step 7: Implement Selected Control(s)

Depending on individual situations, the implemented controls may lower the risk level and may not eliminate the risk completely. This gives room for Residual risks which are kept under close watch and monitoring in the appropriate risk registers

Control Categories

In implementing recommended controls to mitigate risk, an organization should consider technical, management, and operational security controls, or a combination of such controls, to maximize the effectiveness of controls for their IT systems and organization. Security controls, when used appropriately, can prevent, limit, or deter threat-source damage to an organization's mission. The control recommendation process will involve choosing among a combination of technical, management, and operational controls for improving the organization's security posture. The trade-offs that an organization will have to consider are illustrated by viewing the decisions involved in enforcing use of complex user passwords to minimize password guessing and cracking. In this case, a technical control requiring add-on security software may be more complex and expensive than a procedural control, but the technical control is likely to be more effective because the enforcement is automated by the system.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Residual Risks and Updating of Risk Register

As the final output of risk mitigation process, Organizations can analyze the extent of the risk reduction generated by the new or enhanced (recommended) controls in terms of the reduced threat likelihood of occurrence or impact - the two parameters that define the mitigated level of risk to the organizational mission. And Of course, update the risk register either by including list of risks that have least impact when all are compared and/or de-listing the ones which have been triggered and considered of high impact and invariably mitigated.

Implementation of new or enhanced controls can mitigate risk by:

a) Eliminating some of the system's vulnerabilities (flaws and weakness), thereby reducing the number of possible threat-source/vulnerability pairs or outright prevention of exercise of the vulnerabilities
b) Adding a targeted control to reduce the capacity and motivation of a threat-source .For example, a department determines that the cost for installing and maintaining add-on security software for the stand-alone PC that stores its sensitive files is not justifiable, but that administrative and physical controls should be implemented to make physical access to that PC more difficult (e.g., store the PC in a locked room, with the key kept by the manager).
c) Reducing the magnitude of the adverse impact (for example, limiting the extent of vulnerability or modifying the nature of the relationship between the IT system and the organization's mission).

III RISK MITIGATION OPTIONS FOR NEGATIVE AND POSITIVE RISKS RESPECTIVELY

These are various options through which threats can be reduced and where possible, eradicated (in case of negative risks). However, in case of positive risks - opportunities, they can be enhanced or invested upon. Based on these two sides of a risk, its mitigation options are treated in these lights as well

Strategies for Negative Risks or threats: There are three main strategies used to deal with threats that may lead to compromise of data/information integrity and confidentiality by exploiting the vulnerability in the devices; if they occur are:

a) Risk Avoidance: This is used where the risk impact is high. The stakeholders act to eliminate the threats .The most radical avoidance strategy is to shut down the devices
or disconnect them from network. This may prompt the stakeholders to consult the manufacturers for immediate solution, if there is no other alternative.
b) Risk Transfer: Here, the stakeholders shift the impact of the threat to a third party and ownership of the responsibility by use of insurance, warranties, guarantees etc.
c) Risk Mitigate: In this strategy, stakeholders act early to reduce the probability of occurrence or impact of a risk. Thereby making the risk to be within acceptable threshold.
d) Risk Acceptance: This is used for Negative and Positive risks. In this scenario, stakeholders decide to acknowledge the risks and take no action unless the risk occurs. However this strategy provides room for periodic reviews of the threats to ensure that the risk does not change significantly. This also happens for the risks under close monitoring such as those in the risk registers.

Strategies for Positive Risks or Opportunities:

a) Exploit: This is used for risks with positive impacts on the devices where the stakeholders wish to ensure the opportunity is realized. It seeks to eliminate the uncertainty associated with a particular upside risk by ensuring the opportunity definitely happens. For example, engaging a vast expert to administer the devices who ensures that all the devices` full potential are utilized and also embraces trends of new technologies including their upgrades in order to proactively minimized any vulnerability and negative risks.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

b) Enhance: This is used to increase probability and/or positive impacts of an opportunity. Identifying and Maximizing key drivers of this positive-impact risk may increase the probability of their occurrence. For example, Changing/upgrading the software (Operating systems, application etc.) and hardware of a device will definitely increase the throughput and security.

c) Share: Sharing a positive risk involves allocating some or all of the ownership of the opportunity to a third party who is best able to capture the opportunity for the benefit of the stakeholders. For example, forming a risk-sharing partnership, teams or joint ventures can be established with express purpose of taking advantage of the opportunity so that all stakeholders gain from their actions.

d) Accept: Accepting an opportunity is being willing to take advantage of the opportunity if it arises but not practically pursuing it.

The goals and mission of an organization should be considered in selecting any of any of the options. It may not be practical to address all identified risks (low, medium and high), so priority should be given to the risks which adjudged to have potential to cause significant mission impact or harm to the stakeholders and their organizations. Therefore, the best of breed approach is to use appropriate technologies from various vendor security products, along with the appropriate risk mitigation option and other administrative measures which are best practices.

IV. COST-BENEFIT ANALYSIS

To allocate resources and implement cost-effective controls, organizations, after identifying all possible controls and evaluating their feasibility and effectiveness, should conduct a cost-benefit analysis for each recommended control to determine which controls are required and appropriate for their circumstances. High cost of a control does not means most effective and appropriate for a particular risk.

The cost-benefit analysis can be qualitative or quantitative. Its purpose is to demonstrate that the costs of implementing the controls can be justified by the reduction in the level of risk. For example, the organization may not want to spend 1, 000NGN on a control to reduce a 200NGN risk.

A cost-benefit analysis for recommended controls encompasses the following:

a) Determining the impact of implementing the new or enhanced recommended controls
b) Determining the impact of not implementing the new or enhanced recommended controls
c) Compatibility and adaptability of the recommended control to the existing ones
d) Estimating the costs of the implementation. These may include, but are not limited to, the following:

i. Hardware and software purchases or upgrades
ii. Reduced operational effectiveness if system performance or functionality is reduced for increased security
iii. Cost of implementing additional policies and procedures
iv. Cost of hiring additional personnel to implement New policies, procedures, or services, if need be.
v. Training costs
vi. Maintenance costs and /or operational cost if applicable

Assessing the implementation costs and benefits against system and data criticality to determine the importance to the organization of implementing the recommended controls, given their costs and relative impact.

The organization will need to assess the benefits of the controls in terms of maintaining an acceptable mission posture for the organization. Just as there is a cost for implementing a needed control, there is a cost also for not implementing it. By

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

relating the result of not implementing the control to the mission, organizations can determine whether it is feasible to forgo its implementation. At this stage after considering the above caveats, it is recommended that best-and-purpose fit recommended control be implemented.

V. Conclusions

Identifying risks is important but mitigating them is much more important, as ensuring safety of stakeholders` investments and of course, mission protection is one of every organization`s principal targets. Observing and following the steps as earmarked in the process is key to ensure fulfilment of risk mitigation process. In addition, cost-benefit analysis cannot be overlooked as it is a critical component of the entire process. Hence every recommended control or option must be deeply analyzed to ensure fit for purpose and further assurance of targeted results or impact. In all these, there should be consistent updating of stakeholders and sustenance of the communication throughout life span of the organization. This would further enshrine the process among stakeholders and build organization-wide proactive focus on mitigating high and/or critical medium risks.

VI. References

Stoneburner G.,Goguen A. and Feringa A.(2001,July). Risk Management Guide for information Systems. Retrieved September 4,2014 from http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf.

PMI. (2012).A Guide to the Project Management Body of Knowledge (PMBOK) Fifth Edition.

Cooney M.(2012,September 21).10 common mobile Security problems to attack. Retrieved August 22,2014 from http://www.pcworld.com/article/2010278/10-common-mobile-security-problems-to-attack.html

A Guide to the Project Management Body of Knowledge(PMBOK) Fifth Edition.

Siciliano, R. S. (2011). Seven Types of Hacker Motivations. Infosec Island

Risk Management Process Overview. Accessed September 3,2014 from http://technet.microsoft.com/en-us/library/cc535304.aspx

Schneier, B. (2008). Security through Obscurity.

Siciliano, R. S. (2011). Seven Types of Hacker Motivations. Infosec Island

Valsamakis, A. C., (2003). Risk Management.
Heinemann Higher and Further Education (Pty) Ltd. Sandton.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Credit Card Fraud Detection Using Neural Network

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem; and Inyiama, Hycient

[1] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[2] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

The payment card industry has grown rapidly the last few years. Companies and institutions move parts of their business, or the entire business, towards online services providing e-commerce, information and communication services for the purpose of allowing their customers better efficiency and accessibility. Regardless of location, consumers can make the same purchases as they previously did "over the desk". The evolution is a big step forward for the efficiency, accessibility and profitability point of view but it also has some drawbacks. The evolution is accompanied with a greater vulnerability to threats. The problem with making business through the Internet lies in the fact that neither the card nor the cardholder needs to be present at the point-of-sale. It is therefore impossible for the merchant to check whether the customer is the genuine cardholder or not. Payment card fraud has become a serious problem throughout the world. Companies and institutions loose huge amounts annually due to fraud and fraudsters continuously seek new ways to commit illegal actions. The good news is that fraud tends to be perpetrated to certain patterns and that it is possible to detect such patterns, and hence fraud. In this paper we will try to detect fraudulent transaction through the neural network along with the genetic algorithm. As we will see that artificial neural network when trained properly can work as a human brain, though it is impossible for the artificial neural network to imitate the human brain to the extent at which brain work, yet neural network and brain, depend for there working on the neurons, which is the small functional unit in brain as well as ANN. Genetic algorithm are used for making the decision about the network topology, number of hidden layers, number of nodes that will be used in the design of neural network for our problem of credit card fraud detection. For the learning purpose of artificial neural network we will use supervised learning feed forward back propagation algorithm.

1. INTRODUCTION

There are many ways in which fraudsters execute a credit card fraud. As technology changes, so does the technology of fraudsters, and thus the way in which they go about carrying out fraudulent activities. Frauds can be broadly classified into three categories, i.e., traditional card related frauds, merchant related frauds and Internet frauds. The different types of methods for committing credit card frauds are described below.

Merchant Related Frauds

Merchant related frauds are initiated either by owners of the merchant establishment or their employees. The types of frauds initiated by merchants are described below:

i. Merchant Collusion **:** This type of fraud occurs when merchant owners or their employees conspire to commit fraud using the cardholder accounts or by using the personal information. They pass on the information about cardholders to fraudsters.

ii. Triangulation: Triangulation is a type of fraud which is done and operates from a web site. The products or goods are offered at heavily discounted rates and are also shipped before payment. The customer while browse the site and if he likes the product he place the online information such as name, address and valid credit card details to the site. When the fraudsters receive these details, they order goods from a legitimate site using stolen credit card details. The fraudsters then by using the credit card information purchase the products.

Internet Related Frauds

The internet is the base for the fraudsters to make the frauds in the simply and the easiest way. Fraudsters have recently begun to operate on a truly transnational level. With the expansion of trans-border, economic and political spaces, the internet has become a new worlds market, capturing consumers from most countries

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

around the world. The below described are most commonly used techniques in Internet fraud:

i. Site clonin: Site cloning is where fraudsters close an entire site or just the pages from which the customer made a purchase. Customers have no reason to believe they are not dealing with the company that they wished to purchase goods or services from because the pages that they are viewing are identical to those of the real site. The cloned site will receive these details and send the customer a receipt of the transaction through the email just as the real company would do. The consumer suspects nothing, while the fraudsters have all the details they need to commit credit card fraud.

ii. *False merchant sites***:** Some sites often offer a cheap service for the customers. That site requests the customer to fill his complete details such as name and address to access the webpage where the customer gets his required products. Many of these sites claim to be free, but require a valid credit card number to verify an individual s age. These kinds of sites in this way collect as many as credit card details. The sites themselves never charge individuals for the services they provide. The sites are usually part of a larger criminal network that either uses the details it collects to raise revenues or sells valid credit card details to small fraudsters.

iii. *Credit card generators***:** These are the computer programs that generate valid credit card numbers and expiry dates. These generators work by generating lists of credit card account numbers from a single account number. The software works by using the mathematical Luhn algorithm that card issuers use to generate other valid card number combinations. This makes the user to allow to illegally generating as many numbers as he desires, in the form of any of the credit card formats.

FRAUD DETECTION USING NEURAL NETWORK

Although there are several fraud detection technology exist based on Data mining, Knowledge Discovery and Expert System etc. but all these are not capable enough to detect the fraud at the time when fraudulent transaction are in progress due to very less chance of a transaction being fraudulent .It has been seen that Credit card fraud detection has two highly peculiar characteristics The first one is obviously the very limited time span in which the acceptance or rejection decision has to be made.

The second one is the huge amount of credit card operations that have to be processed at a given time. To just give a medium size example, millions of Visa card operations take place in a given day, 98% of them being handled on line. Of course, just very few will be fraudulent (otherwise, the entire industry would have soon ended up being out of businesses), but this just means that the haystack where these needles are to be found is simply enormous.

Working principal (Pattern Recognition)

Neural network based fraud detection is based totally on the human brain working principal. Neural network technology has made a computer capable of think. As human brain learn through past experience and use its knowledge or experience in making the decision in daily life problem the same technique is applied with the credit card fraud detection technology. When a particular consumer uses its credit card , there is a fix pattern of credit card use , made by the way consumer uses its credit card.

Using the last one or two year data neural network is train about the particular pattern of using a credit card by a particular consumer. As shown in the figure the neural network are train on information regarding to various categories about the card holder such as occupation of the card holder, income, occupation may fall in one category, while in another category information about the large amount of purchased are placed, these information include the number of large purchase, frequencies of large purchase, location where these kind of purchase are take place etc. within a fixed time period.

In spite of pattern of credit card use neural network are also trained about the various credit card fraud face by a particular bank previously. Based on the pattern of uses of credit card, neural network make use of prediction algorithm on these pattern data to classify that weather a particular transaction is fraudulent or genuine.

When credit card is being used by unauthorized user the neural network based fraud detection system check for the pattern used by the fraudster and matches with the pattern of the original card holder on which the neural network has been trained, if the pattern matches the neural network declare the transaction ok

When a transaction arrives for authorization, it is characterized by a stream of authorization data fields that carry information identifying the cardholder (account number) and characteristics of the transaction (e.g., amount, merchant code). There are additional data fields that can be taken in a feed from the authorization system (e.g., time of day). In most cases, banks do not archive logs of their authorization files. Only transactions that are forwarded by the merchant for settlement are archived by the bank's credit card processing system. Thus, a data set of transactions was composed from an extract of data stored in Bank's settlement file. In this extract, only that authorization information that was archived to the settlement file was available for model development.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Fraud Detection

Matching the pattern does not mean that the transaction should exactly match with the pattern rather the neural network see to what extent there exist difference if the transaction is near by the pattern then the transaction is ok otherwise if there is a big difference then the chance of being a transaction illegal increase and the neural network declare the transaction a fault transaction.

The neural network is design to produce output in real value between 0 and 1 .If the neural network produce output that is below .6 or .7 then the transaction is ok and if the output is above .7 then the chance of being a transaction illegal increase.

There are some occasions when the transaction made by a legal user is of a quite different and there are also possibilities that the illegal person made use of card that fit into the pattern for what the neural network is trained. Although it is rare, yet If the legal user can't complete a transaction due to these limitation then it is not much about to worry But what about the illegal person who is making use of card , hare also work human tendency to some extent when a illegal person gets a credit card he is not going to make use of this card again and again by making number of small transaction rather he will try to made as large purchase as possible and as quickly that may totally mismatch with the pattern for what the neural network is trained.

TRANSACTION FRAUD SCORER

The neural network used in this fraud detection a three-layer, feed-forward network that use two training passes through the data set. The fast training pass involves a process of prototype cell commitment in which exemplars from the training set are stored in the weights between the first and second (middle) layer cells of the network. A final training pass determines local a posteriori probabilities associated with each of these prototype cells. P-RCE training is not subject to problems of convergence that can afflict gradient-descent training algorithms. The P-RCE network and networks like it have been applied to a variety of pattern recognition problems both within and beyond the field of financial services, from character recognition to mortgage underwriting and risk assessment layer consisted of a single cell that outputs a numeric response that can be considered as a "fraud score". This is analogous to credit scoring systems that produce a score, as opposed to a strict probability. The objective of the neural network training process is to arrive at a trained network that produces a fraud score that gives the best ranking of the credit card transactions. If the ranking were perfect, all of the high scoring transactions down to some threshold would be fraud; below this threshold, only good transactions would be ranked.

However, perfect separation of frauds from goods is not possible due to the inherently non-separable nature of the fraud and good distributions in the selected pattern recognition Space.

Final evaluation of the trained network can be done on the Blind Test data set. The Blind Test data represented an unsampled set of all Banks' transactions during last few months.

Learning Algorithm (Feed Forward Back Propagation)

The back propagation learning rule is a standard learning technique. It performs a gradient descent in the error/ weights space. To improve the efficiency, a momentum term is introduced, which moves the correction of the weights in the direction compliant with the last weight correction?

It is a multi-layer feed forward network that is trained by supervised learning.

Supervised learning means that the network is repeatedly presented with input/output pairs (I,O) provided by a supervisor, where O is the output the network should produce when presented with input I. These input/output pairs specify the activation patterns of the input and output layer. The network has to find an internal representation that result in the wanted input/output behavior. To achieve this, back propagation uses a two-phase propagates-adapt cycle.

i. First Phase: In the first phase the input is presented to the network and the activation of each of the nodes (processing elements) of the input layer is propagated to the hidden layer, where each node sums its input and propagates its calculated output to the next layer. The nodes in the output layer calculate their activations in the same way as the nodes in the hidden layer.

ii. Second Phase: In the second phase, the output of the network is compared with the desired output given by the supervisor and for each output node the error is calculated. Then the error signals are transmitted to the hidden layer where for each node its contribution to the total error is calculated. Based on the error signals received, connection weights are then *adapted* by each node to cause the network to converge toward a state that allows all the training patterns (input/output pairs) to be encoded.

PROBLEM WITH THE TRAINING OF NEURAL NETWORK

Problem with neural networks is that a number of parameter has to be set before any training can begin. However, there are no clear rules how to set these parameters. Yet these parameters determine the success of the training. In the most general case, neural networks consist of an (often very high) number of neurons, each of which has a number of inputs which are mapped via a relatively simple function to its output.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Networks differ in the way their neurons are interconnected (topology), in the way the output of a neuron determined out of its inputs (propagation function) and in their temporal behavior (synchronous, asynchronous or continuous).

The topology of a network has a large influence on the performance of that network but, so far, no method exists to determine the optimal topology for a given problem because of the high complexity of large networks. the choice of the basic parameter (network topology, learning rate, initial weights) often already determines the success of the training process. The selection of these parameters follow in practical use rules of thumb, but their value is at most arguable.

Genetic Algorithms Overview

The biological metaphor for genetic algorithms is the evolution of the species by survival of the fittest, as described by Charles Darwin. In a population of animals or plants, a new individual is generated by the crossover of the genetic information of two parents.

The genetic information for the construction of the individual is stored in the DNA. The human DNA genome consists of 46 chromosomes, which are strings of four different bases, abbreviated A, T, G and C. A triple of bases is translated into one of 20 amino acids or a "start protein building" or "stop protein building" signal. In total, there are about three billion nucleotides. These can be structured in genes, which carry one or more pieces information about the construction of the individual. However, it is estimated that only 3% of the genes carry meaningful information, the vast majority of genes - the "junk" genes - is not used.

The genetic information itself, the genome, is called the *genotype* of the individual. The result, the individual, is called *phenotype***.** The same genotype may result in different phenotypes. Twins illustrate this quite well.

Genetic algorithms are algorithms for optimization and machine learning based loosely on several features of biological evolution. They require five components:

i. A way of encoding solutions to the problem on chromosomes.

ii. An evaluation function which returns a rating for each chromosome given to it

Iii.A way of initializing the population of chromosomes.

iv. Operators that may be applied to parents when they reproduce to alter their genetic composition. Standard operators are mutation and crossover Parameter settings for the algorithm, the operators, and so forth.

GENETIC ALGORITHM ALONG WITH NEURAL NETWORK

(GANN) By combining genetic algorithms with neural networks (GANN), the genetic algorithm is used to find these parameters. The inspiration for this idea comes from nature:

In real life, the success of an individual is not only determined by his knowledge and skills, which he gained through experience (the neural network training), it also depends on his genetic heritage (set by the genetic algorithm). One might say, GANN applies a natural algorithm that proved to be very successful on this planet: It created human intelligence from scratch. The main question is how exactly GA and NN can be combined, i.e. especially how the neural network should be represented to get good results from the genetic algorithm

Information about the neural network is encoded in the genome of the genetic algorithm. At the beginning, a number of random individuals are generated. The parameter strings have to be evaluated, which means a neural network has to be designed according to the genome information. Its performance can be determined after training with back-propagation. Some GANN strategies rely only on the GA to find an optimal network; in these, no training. take place. Then, they are evaluated and ranked. The fitness evaluation may take more into consideration than only the performance of the individual.

Principle Structure of GA and GANN System

Individual's version and the network pruning algorithm. The first uses just the weight-encoding bits, the second merely the index-bit. For the later, the weight values of an already generated optimal network are used, the goal is to find a minimal network with good performance. Of course, the number of weights pruned has to be considered in the fitness function. GENITOR requires that a basic (maximal) architecture has to be designed for each problem. The resulting encoding format is a bit-string of fixed length.

The standard GA has no difficulties to deal with this genome. Since crossover can take place at any place of the bit string, a child may have a different weight value than either one of the parents. So, topology and weight values are optimized at the same time. Whitley reports that GENITOR tends to converge to a single solution, the diversity is reduced fast. It seems to be a good "genetic hill-climber". The approach was applied to simple Boolean functions.

CONCLUSION

In this paper we saw different technique that is being used to execute credit card fraud how credit card fraud impact on the financial institution as well as merchant and customer, fraud detection technique used by VISA and MasterCard. Neural network is a latest technique that is being used in different areas due to its powerful capabilities of learning and predicting.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In this thesis we try to use this capability of neural network in the area of credit card fraud detection as we know that Back propagation Network is the most popular learning algorithm to train the neural network so in this paper BPN is used for training purpose and then in order to choose those parameter (weight, network type, number of layer, number of node e.t.c) that play an important role to perform neural network as accurately as possible, we use genetic algorithm, and using this combined Genetic Algorithm and Neural Network (GANN) we try to detect the credit card fraud successfully. The idea of combining Neural Network and genetic Algorithm come from the fact that if a person is inherently very talented and he is trained properly then chances of individual of success is very high.

REFERENCES

Abidi, S. S. R. (2001). Knowledge management in healthcare: towards 'knowledge-driven' decision-support services. *International Journal of Medical Informatics*, *63*(1-2), 5–18.

Ahmad, R., Kausar, A. R., & David, P. (Writer) (2007). *The social management of embodied knowledge in a knowledge community*.

Anderson, R. A., & McDaniel, R. R. (2000). Managing health care organizations: where profes-sionalism meets complexity science. *Health Care Management Review*, *25*(1), 83–92.

Andreas, R. (2005). Three-dozen knowledge-sharing barriers managers must consider. *Jour-nal of Knowledge Management*, *9*(3), 18–35.

Andreas, R., & Nicholas, L. (2006). Knowledge management in the public sector: stakeholder partnerships in the public policy development. *Journal of Knowledge Management*, *10*(3), 24.

Ansell, C. (2007). Fostering Innovation and Col-laboration. *Medical Device Technology*, *18*(1), 52.

Bali, R. K., & Dwivedi, A. N. (Eds.). (2007). *Healthcare Knowledge Management*: Springer.

Batalden, P., & Splaine, M. (2002). What will it take to lead the continual improvement and in-novation of health care in the twenty-first century? *Quality Management in Health Care*, *11*(1), 45.

Bate, S. P., & Robert, G. (2002). Knowledge management and communities of practice in the private sector: lessons for modernizing the National Health Service in England and Wales. *Public Administration*, *80*(4), 643–663.

Bates, D. W., Spell, N., Cullen, D. J., Burdick, E., Laird, N., & Petersen, L. A. (1997). The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. *Journal of the American Medical Association*, *277*(4), 307–311. doi:10.1001/jama.277.4.307

Bontis, N. (1996). There's a price on your head: Managing Intellectual Capital Strategically. *Business Quarterly*, (Summer): 40–47.

Bontis, N. (2001). Assessing knowledge assets: a review of models used to measure intellectual capital. *International Journal of Management Reviews*, *3*(1), 41–60.

Brian, D. P. (2006). *Final Report of the Federal Advisor on Wait Times*. Ottawa, Canada: Health Canada.

Buchan, I. E., & Hanka, R. (1997). Exchanging clinical knowledge via Internet. *International Journal of Medical Informatics*, *47*(1-2), 39–41.

Burnett, S. M., Williams, D. A., & Webster, L. (2005). Knowledge support for interdisciplin-ary models of healthcare delivery: a study of knowledge needs and roles in managed clinical networks. *Health Informatics Journal*, *11*(2), 146–160.

Caldwell, D. F., Chatman, J., O'Reilly, C. A., III, Ormiston, M., & Lapiz, M. (2008). Implementing strategic change in a health care system: the impor-tance of leadership and change readiness. *Health care management review*, *33*(2), 124(110).

Canadian Health Services Research Foundation. (2003). *The theory and practice of knowledge*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REINFORCEMENT LEARNING A TOOL FOR FILTERING PERSONALIZED WEB DOCUMENT

Nwagu, Chikezie Kenneth, Omankwu, Obinnaya Chinecherem; and Inyiama, Hycient

[1] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[2] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

Document filtering is increasingly deployed in Web environments to reduce information over-load of users. We formulate online information filtering as a reinforcement learning problem, i.e. TD(0). The goal is to learn user profiles that best represent his information needs and thus maximize the expected value of user relevance feedback. A method is then presented that acquires reinforcement signals automatically by estimating user's implicit feedback from direct observations of browsing behaviors. This “learning by observation” approach is contrasted with conventional relevance feed-back methods which require explicit user feedbacks. Field tests have been performed which involved 10 users reading a total of 18,750 HTML documents during 45 days. Compared to the existing document filtering techniques, the proposed learning method showed superior performance in information quality and adaptation speed to user preferences in online filtering.

Keywords
Web based, Document, Reinforcement learning

1. INTRODUCTION

With the rapid progress of computer technology in recent years, electronic information has been explosively increased. This trend is especially remarkable on the Web. As the availability of the information increases, the need for finding more relevant information on the Web is growing [Belkin and Croft, 1996]. Currently, there are two major ways of accessing information on the Web. One is to use Web index services such as AltaVista, Yahoo, and Excite. The other is to manually follow or browse the hyperlinks of the documents by a user himself. However, these methods have some drawbacks. Since Web-index services are based on general purpose indexing methods, much of the retrieval results may be irrelevant to user's interests. In addition, manual browsing involves much time and efforts. High-quality information services require to capture the personal interests of individual users during the interaction with the information retrieval systems.

Several methods have been proposed to reflect user preferences. A classical approach is the Rocchio method [Rocchio, 1971] and its variants. This is a batch algorithm that modifies the original query vector by the vectors of the relevant and irrelevant documents. However, the batch algorithms tend to put large demands on memory and are slow in adaptation, thus not well suited to on-line applications. Recently, several on-line learning algorithms have been used for information retrieval and filtering. These include the Widrow-Hoff rule [Lewis et al., 1996] and the exponentiated gradient algorithm [Callan, 1998]. These algorithms learn training examples one at a time and thus more appropriate for learning in online fashion. However, all these methods have a drawback that the user has to provide explicit relevance feedback for the system to learn. Since providing relevance feedbacks is a tedious process and users may be unwilling to provide them, the learning capability of the filtering systems may be severely limited.

In this paper, we present a personalized information filtering method that learns user's interests by observing his or her behaviors during the interaction with the system. First, the system is trained on the explicit feed-back from the user. After this learning phase, the system estimates the relevance feedback implicitly based on the observations of user actions. This information is used to modify the user profiles. We regard filtering as a goal-directed learning process based on interactions with the environment. The objective is to maximize the expected value of the cumulative relevance feedback it receives in the long run from the user. This process is formulated as TD(0) learning, a general form of reinforcement learning [Sutton and Barto, 1998]. In this formulation, filtering is viewed as an interactive process which involves a generate-and-test method whereby the agent try actions,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

observe the outcomes, and selectively retain those that are the most effective. The advantage of TD(0) over other reinforcement learning methods is that it can learn without excessive delay of rewards. This is an important property in real-time interactions with the user in Web browsing environments. Additional feature of our approach is that it is learning by experimentation, in contrast to learning by instruction as adopted in most supervised learning methods. The method was implemented as WAIR (Web Agents for Information Retrieval), a platform for Web-based personalized information filtering services [Seo and Zhang, 2000].

Personalized Filtering as Reinforcement Learning

Information Filtering in WAIR

WAIR (Web Agents for Information Retrieval) was originally designed as a platform for the development of personalized information services on the Web. WAIR consists of three agents: an interface agent, a re-trieval agent, and a filtering agent. The interaction between the agents is illustrated in Figure 1. The overall procedure is summarized in Figure 2.

Initially, the user provides the system with a profile (Step 1). Typically, the initial profile consists of a few keywords. Then, the retrieval agent constructs a query using the profile and get N URLs (Step 2). Existing Web search engines are used to obtain the relevant URLs. The documents for the URLs are then retrieved and preprocessed, and their relevance values are estimated. The N documents are ranked, and M of them are filtered and presented to the user (Step 3). To balance exploration and exploitation, WAIR chooses the highest-ranked documents most of the time, but occasionally (with probability) it filters lower-ranked documents.

The interface agent observes user behavior and measures user feedback (Step 4). Two different types of user feedbacks are distinguished in WAIR. One is the “explicit” feedback in the form of scalar values to evaluate the relevance of the documents. This is provided by the user during the initial learning phase. A second type of feedback is the “implicit” feedback. This is not provided by the user, but estimated by the interface agent in WAIR .

That is, the users read filtered HTML documents by performing normal browsing behaviors, such as scrolling thumb up and down, bookmarking an URL, following the hyperlinks in the filtered document, and the WAIR infers from the behaviors how much the user was interested in each filtered document with a multi-layer neural network. This process is described in detail in the next section. The feedback information is then used to update the user profile (Step 5). Basically, this consists of inserting new terms, removing existing terms, and adjusting term weights of profile terms using the terms in the

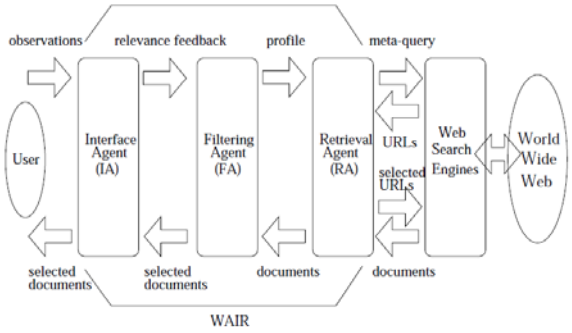


Figure 1: System architecture of WAIR

1. Get the initial profile from the user. Set $t \leftarrow 0$.
2. (Retrieval) Generate a query from the profile to retrieve N URLs.
3. (Filtering) Evaluate the relevance of documents.
   Rank the N documents and present M of them to the user.
4. (Interface) Get the feedback by observing user behavior.
5. (Learning) Update the user profile.
6. Set $t \leftarrow t + 1$. Go to step 2.

Figure 2: The overall procedure of WAIR

relevant/irrelevant documents. Then, the revised profile is used to get new documents by going to the retrieval step. Note that, the user provides only an initial query and then WAIR automatically retrieves and filters documents by observing user behaviors implicitly.

Filtering as Reinforcement Learning

The task of information filtering in WAIR is formulated as a reinforcement learning problem. Reinforcement learning is about learning from interaction how to behave in order to achieve a goal. The reinforcement learning agent and its environment interact over a sequence of discrete time steps. The actions are the choices made by the agent. The states are the basis for making the choices.

The rewards are the basis for evaluating choices. In WAIR, actions are defined as the decision-making as to whether to present the document to the user or not. States are defined as the pairs of the profile and the document to be filtered.

The policy is a stochastic rule by which the agent selects actions as a function of states. Formally, a policy is a mapping from each state

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

s and action as to the probability (s; a) of taking action a when in states. We use an -greedy policy for choosing an action given a state. That is, most of the time WAIR chooses the highest-ranked documents, but with probability, it chooses lower-ranked documents too. The rationale behind this policy is that it combines exploitation and exploration of search behavior. The selection of documents with the highest relevance value corresponds to exploitation of known information, while selecting random documents encourages exploration of unknown regions to find interesting documents which are unexpected by the user. An advantage of the -greedy method is that, in the limit as the number of actions increases, the probability of selecting the optimal action converges to greater than 1 , , i.e., to near certainty [Sutton and Barto, 1998].

The filtering agent's objective is to maximize the amount of reward it receives over time. The return is the function of future rewards that the agent seeks to maximize. Value functions of a policy assign to each state, or state-action pair, the expected return from that state, or state-action pair, the largest expected return achievable by any policy. The agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. In particular, it chooses action a_t to maximize the expected discounted return:

$$\begin{aligned} R_t &= r_{t+1} + \gamma r_{t+2} + +\gamma^2 r_{t+3} \cdots \\ &= \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \end{aligned}$$

where γ is a parameter, $0 \leq \gamma \leq 1$, called the discount rate.

To make decisions on whether or not filter the documents, it is necessary to estimate value functions, i.e., functions of states that estimate how good it is to be in a given state. The notion of how good here is defined in terms of future rewards that can be expected, i.e. in terms of expected return. Value functions are defined with respect to particular policies. Informally, the value of a states under a policy , denoted V (s), is the expected return when starting in s and following thereafter. We can define V (s) as

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\{R_t \mid s_t = s\} \\ &= E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s \right\} \\ &= E_{\pi}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,\middle|\, s_t = s \right\} \\ &= E_{\pi}\{ r_{t+1} + \gamma V^{\pi}(s_{t+1}) \mid s_t = s \}, \end{aligned}$$

Learning Profiles from Implicit Feedbacks

In this section, we first describe the retrieval of documents in WAIR. Then, the procedures for estimating user feedbacks and updating user profiles are described.

Document Retrieval

The task of the retrieval agent is to get a collection of candidate HTML documents to be filtered. The retrieved documents undergo preprocessing. We use standard term-indexing techniques, such as removing stop-words and stemming [Frakes and Baeza-Yates, 1992]. Formally, a document is represented as a term vector

$$\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{i,k}, ..., x_{i,d}),$$

where $x_{i;k}$ is the numeric value that term k takes on for document i, d is the number of terms used for document representation. In this work, we assume that $x_{i;k}$ represents the normalized term frequency, i.e. $x_{i;k}$ is proportional to the number of term k appearing in document i and k x_i k= 1. This is contrasted with the usual *tf idf* (term frequency inverse document frequency) [Salton, 1989] based indexing method in conventional information retrieval. We use only tf information because we focus on information filtering from a stream of Web documents. In contrast to the conventional information retrieval environments where the collection of documents is static over a long period of time, our situation addresses a dynamically changing environment. In this dynamic environment, the inverse document frequency (which is computed with respect to a static collection of documents) is not significant.

The ultimate goal of WAIR is to filter documents that best reflect user's preferences. This is done by learn-ing the profiles of users. A user profile consists of one or more topics. Topics represent user's information needs. In this section, we assume for simplicity that a profile consists of a single topic. The method can readily be generalized to multiple topics for a user by maintaining multiple profiles. Formally, the profile p is represented as a weight vector

$$\mathbf{w}_p = (w_{p,1}, w_{p,2}, ..., w_{p,k}, ..., w_{p,d}),$$

where $w_{p;k}$ is the weight of the kth term in the profile and k w_p k= 1. d is the number of terms used for describing the profiles. Formally, it is the same as the number of terms for representing documents. In WAIR, however, the maximum number of non-zero terms in the profile is limited to m < d. This is useful for concise description of user interests. Initially, the profile w_p contains only

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

a small number of non-zero terms that are contained in the original user query. The subsequent retrieval and user-feedback process expands and updates the number and weights of the profile terms, as described below.

WAIR searches the Web-documents by using existing Web-index services, i.e. AltaVista, Excite, and Lycos. That is, it formulates a query q_p that is forwarded to one or more Web search engines. Queries are constructed by choosing terms from the profile based on an -greedy selection method. The retrieval agent then selects *N* URLs from different engines and ranks them. The rank of document i for profile p is based on its similarity (or relevance) to the profile and computed as the inner product:

$$V(s_i) = \mathbf{w}_p \cdot \mathbf{x}_i = \sum_{k=1}^{d} w_{p,k} x_{i,k},$$

where $w_{p;k}$ and $x_{i;k}$ are the kth terms in profile p and document, respectively. The candidate documents are then sorted in descending order of $V_i(s_i)$, and M of them are presented to the user. Note that since the term vectors are normalized to $w_p = 1$ and $x_i = 1$, the relevance value is equivalent to the cosine correlation, i.e.

$$V(s_i) = \frac{\mathbf{w}_p \cdot \mathbf{x}_i}{\| \mathbf{w}_p \| \| \mathbf{x}_i \|}$$

$$\text{where } \| \mathbf{x}_i \| = \sqrt{\sum_{k=1}^{d} x_{i,k}^2}.$$

Conclusions

In this paper, we formulated the problem of information filtering as a TD(0) reinforcement learning problem, and presented a personalized Web-document filtering system that learns to follow user preferences from observations of his behaviors on the presented documents. A practical method was described that estimates the user's relevance feedback from user behaviors such as reading time, bookmarking, scrolling, and link-following actions.

Our experimental evidence from a field test on a group of users supports that the proposed method effectively adapts to the user's specific interests. This confirms that “learning from shoulders of the user” through self-generated reinforcement signals can significantly improve the performance of information filtering systems. In a series of short-term filtering environments, WAIR achieved superior performance when compared to the conventional feedback methods, including Rocchio, WH, and EG. In terms of adaptation speed, the proposed method converged to the user's specific interest faster than existing relevance feedback methods.

Our work has focused on personalizing information filtering based on existing Web-index services, i.e. AltaVista, Excite, and Lycos. Through the use of learning-based personalization techniques, WAIR could improve the quality of information service of the existing Web search engines. Since every search engine has its strengths and weaknesses, the meta-search approach of WAIR combines the strengths of different search engines while reducing their weaknesses. For the convenience of implementation, we used the conventional search engines directly. Using meta-search engines would further increase the final performance. Similar idea can be used to improve the quality of other Web information service systems.

The online nature of reinforcement learning makes it possible to approximate optimal action policies in ways that put more effort into learning to make good decisions for frequently encountered states, at the expense of less effort for infrequently encountered states.

This is the key property that distinguishes reinforcement learning from other relevance feedback methods based on supervised learning. Our experimental result confirms this view: information filtering is dictated by online adaptation based on a small number of documents. The reinforcement learning formulation gave more emphasis on decision making as to filtering the documents rather than just to learn the mappings or profiles. This resulted in better performance than simple supervised learning methods in the dynamic environments. Our work suggests that reinforcement learning can provides a better framework for personalization of information service in the Web environments than conventional supervised learning formulation.

In spite of our success in learning the user preferences in the WAIR system, it should be mentioned that the success comes in part from the environments where we made our experiments.

One is that the topics used for experiments were usually scientific and thus the filtered documents contained relatively less-ambiguous terms than those that might be contained in other usual Web documents. Another reason might be that the duration of our experiments were not very long during which the user interests did not change very much. The adaptation to user's interests during a longer period of time in a more dynamic environment should still be tested. From a more practical point of view, the response time is a crucial factor in the information retrieval and filtering.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

However, our focus in this paper was confined to the relevance feedback. Learning from users to minimize their response time is one of our research topics in the future.

REFERENCES:

Belkin, N.J. and Croft, W.B. 1992. Information filtering and information retrieval: Two sides of the same coin, *Communications of the ACM*, 35(12):29-38.

Boyan J., Freitag D., and Joachims T. 1996. A machine learning architecture for opti-mizing Web search engine, In *Proc. AAAI Workshop on Internet-Based Information Systems*, pp. 324-335.

Callan J. 1998. Learning while filtering documents, In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR-98), pp. 224-231.

Falk, A. and Josson, I.M. 1996. PAWS: An agent for WWW-retrieval and filtering, In *Proc. Practical Application of Intelligent Agents and Multi-agents Technology* (PAAM-96), pp. 169-179.

Frakes, W.B. and Baeza-Yates, R. 1992. Stemming algorithms, In *Informa-tion Retrieval: Data Structures and Algorithms*, pp. 131-160, Prentice Hall.

Hirashima, T., Matsuda, N., Nomoto, T., and Toyoda, J. 1998. Context-sensitive filtering for browsing in hypertext, In *Proc. Int. Conf. on Intelligent User Interfaces* (IUI-98), pp. 119-126.

Joachims, T., Freitag D., and Mitchell, T.M. 1997. WebWatcher: A tour guide for the World Wide Web, In *Proc. Int. Joint Conf. on Artificial Intelligence* (IJCAI-97), pp. 770-777.

Kamba, T., Sakagami, H., and Koseki, Y. 1997. ANATAGONOMY: A personalized newspaper on the World Wide Web, *Int. Jor. of Human-Computer Studies*, Vol. 46, pp. 789-803.

Kindo, T., Yoshida, H., Morimoto, T., and Watanabe, T. 1997. Adaptive personal in-formation filtering system that organizes personal profiles automatically, In *the Proc. Int. Joint Conf. on Artificial Intelligence* (IJCAI-97), pp. 716-721.

Lashkari, Y., Metral, M., and Maes, P., 1994. Collaborative interface agents, In *Proc.of the Twelfth National Conf. on Artificial Intelligence* , pp.444-450.

Lewis, D.D., Schapire, R.E., Callan, J.P., and Papka, R. 1996. Training algorithms for linear text classifiers, In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR-96), pp. 298-306.

Lieberman, H. 1995. Letizia: An agent that assists Web browsing, In *Proc. Int. Joint Conf. on Artificial Intelligence* (IJCAI-95), pp. 475-480.

Maes, P. 1994. Agents that reduce work and information overload, *Communications of the ACM*, 37(7):31-40.

Mitchell, T.M. 1997. *Machine Learning*, McGraw-Hill.

Morita, M. and Shinoda, Y. 1994. Information filtering based on user behav-ior analysis and best match text retrieval, In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval* (SIGIR-94), pp. 272-281.

Pazzani, M. and Billsus, D. 1997. Learning and revising user profiles: the iden-tification of interesting Web sites, *Machine Learning*, 27:313-331.

Rocchio, J.J. 1971. Relevance feedback in information retrieval, In *The SMART Retrieval System*, Prentice Hall, pp. 313-323.

Salton, G. 1989. *Automatic Text Processing*, Addison Wesley.

Salton, G. and Buckley, C. 1990. Improving retrieval performance by relevance feedback, *Journal of American Society for Information Science*, 41:288-297.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Influence of Distributed Generation on the Protection Scheme of Local Distribution System

Usman Yousaf[1], Dr. Azzam ul Asar[2], Alishpa Khattak[3]
Department of Electrical Engineering, CECOS University, Peshawar, Pakistan
emails: usm_usf@hotmail.com[1], dean@cecos.edu.pk[2], alishpakhattak92@gmail.com[3]

***Abstract*— The integration of Distributed Generation (DG) into the distribution system is considered as an achievement made in the field of power system. With such penetration of DG requires assessment of protection schemes used in traditional distribution system. Short circuit studies are needed to be performed for determining an adequate protection scheme for such an integrated distribution system. This paper presents short circuit analysis of 132 KV Dargai Grid Station (GS) Pakistan incompliance to IEC 60909.The Distribution system is modeled in Electrical Transients and Analysis Program (ETAP) software and comparative fault analysis has been performed with and without DG. The fault location is made fixed while DG location is varied. It is found that there is significant increase in fault current with the DG Penetration and the fault current depends upon total feeder length, distance of fault location from the DG and Grid and extent of flowing current.**

***Index Terms*—Distributed Generation, ETAP, IEC 60909, Short Circuit Analysis, Distribution System**

I. INTRODUCTION

The yearly growing electrical energy demand has increased the penetration of DG significantly in to the distribution network. Distribution system is the link between the end user and the utility system [1]. Various benefits are provided to the utility and the consumer by interconnecting DG to an existing distribution system. DG provides an enhanced power quality, higher reliability of the distribution system and peak shaving. However, power system protection being one of the major issue several technical problems are associated with the integration of DG into existing distribution system. The radial power flow is lost and the fault level of the system is increased due to the incorporation of DG [2].

Short circuit studies are one of the most important tasks in power system analysis. According to IEC 60909 short circuit is the accidental or intentional conductive path between two or More conductive parts forcing the electric potential difference between these conductive parts become zero [3].

Short circuit currents produce powerful magnetic forces and intense heat in the power system, which can result in considerable damage to the power system protective equipment. As the breaking capacity of circuit breakers is described by the initial symmetrical fault current flows through the system .when fault occurs, these values of short-circuit currents must be determined to ensure that the short-circuit ratings of all equipment are adequate to sustain the currents available at their locations [3].

This paper aims to verify the effect of DG on the fault current contribution and also investigates viable location for DG coupling by considering three scenarios based on peak load, feeder length and fault location from DG of the distribution system using ETAP software [4], [5].

II. IEC 60909 SHORT CIRCUIT ANALYSIS

IEC 60909 Short Circuit Currents in Three Phase System describes an internationally accepted method for the calculation of fault currents. In applying the standard, two levels of fault based on voltage factor are typically calculated [6],[7].

- The maximum current which causes the maximum thermal and electromagnetic (Mechanical) effects on equipment and is used to determine the equipment rating.
- The minimum current which is used for the setting of protective devices such as relay settings and coordinated relay operation.

Depending on the position within the cycle at which the fault forms, a dc offset will be present, decaying overtime to zero. This creates an initial symmetrical short circuit I''_k, which will decay over time to the steady state short circuitI_k. [3]

A. Initial AC symmetrical short circuit fault current I''_k

The Maximum initial short circuit occurs for a system when three phase fault develop. This current is the root mean square value of the initial component of the short circuit current, which can be calculated by eq 1.[6]

$$I''_K = \frac{C.U_n}{\sqrt{3}\cdot Z_K} = \frac{C.U_n}{\sqrt{3}.\sqrt{R^2 \pm X^2}} \qquad (1)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Where
U_n=Nominal voltage
Z_K=equivalent short circuit impedance at the fault location
C=voltage correction factor

The "C" factor or voltage factor is the ratio of equivalent voltage to the nominal voltage, and required to account for variation due to time and place, transformer taps, static load and capacitance, generator and motor sub transient behavior. [3]

B. *Peak Short Circuit current I_P*

It is the maximum momentary value of the short circuit current. It is only calculated for maximum short circuit current and can be calculated by eq.2

$$\boldsymbol{I_P=\sqrt{2}\times k\times I_K''} \qquad (2)$$

"K" is the function of system R/X ratio at the fault location and can be calculated by eq.3

$$\mathbf{k=1.02+0.98.}e^{\frac{-3R}{X}} \qquad (3)$$

At a fault location the F, the total amount of peak short circuit current is the absolute value of all partial short circuit currents as shown by eq.4. When the R/X ratio remains less than 0.3 at all branches, the R/X ratio of equivalent impedance at the fault location can be used for calculation of k [6].

$$\boldsymbol{i_p=\sum_i i_{pi}} \qquad (4)$$

The system R/X ratio depends on the method selected for calculation. Method A is for uniform
X/R ratio. Method B is for meshed networks and Method C is for non-meshed networks. [3]

C. *Steady State Short Circuit Current I_K*

It is the value of short circuit current when several cycles have been passed. For calculation of maximum steady short circuit current the synchronous generator excitation is kept at maximum.

III. System Description And Simulation Model

For determining actual performance of the power system, the proper mathematical model, accurate parameters of the power network, the generators, transformers and actual loads have to be identified [3]. Actual data related to the transformers, generators, load, and electrical parameters is collected from the power houses and the grid. Fig. 1 shows single line diagram of 132 kV GS Dargai Pakistan simulated in ETAP software [8], [9].

The Dargai Grid station is connected to 81 MW Malakand III Hydro Power Complex which has three 27.2 MW Generator Units transmitting power by Two 132KV outgoing transmission lines. The Grid is also connected by single 132 KV incoming Transmission line to 20 MW Dargai Power House which consists of two 10 MW Generator Units. In Malakand III Hydro Power Complex three 32 MVA Power Transformers step up 11 kV Generated voltage to 132 KV and in Dragai power House two 15 MVA Power Transformers steps up 11 KV generated Voltage to 132 KV, Two 20/26 MVA Distribution Transformers are installed at 132 KV Dargai Grid station which steps down the incoming 132 KV into 11 KV.

Generating units in study are represented by detailed model, with transient and sub-transient circuits on both the direct and quadrate axes been considered, as it describes all possible contribution to the short circuit current.

The Grid station is also connected by two Transmission lines to Chakdara and Mardan Power grids having 568.18 MVA_{SC} and 793.65 MVA_{SC} capacity respectively.

Twelve 11 KV local radial distribution feeders are emanating from the GS to the consumers. The total real time maximum current on the Grid station is 3420 A.

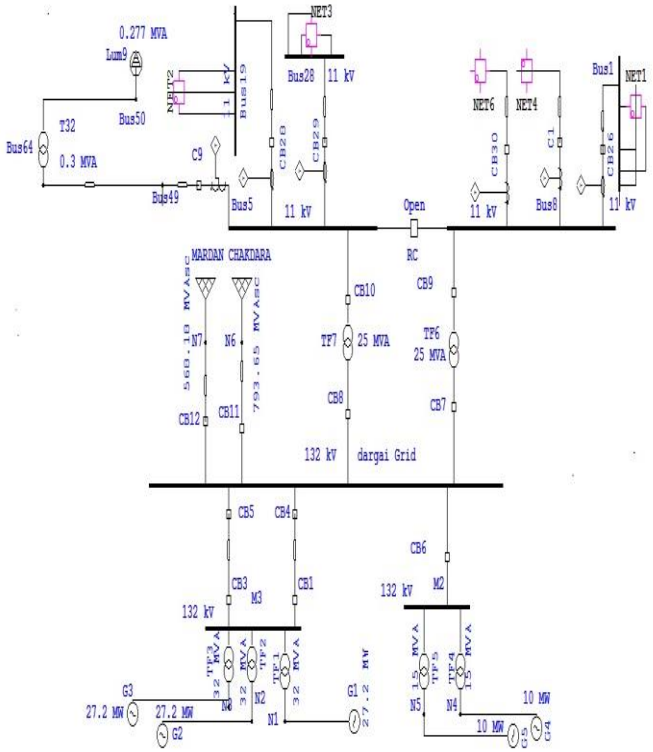


Fig.1 modeled single line diagram of 132 KV grid station Dargai with DG

IV. Case Analysis And Simulation Results

In this paper short circuit analysis is carried out on four feeders, firstly fault currents are obtained without any DG penetration while in other four cases different locations of DG are considered based on various feeder parameters so as to investigate the influence of DG in contribution of short circuit current under fault condition, the case wise simulation results are described below [10].

A. *Fault Analysis without DG.*

Case 1:

The system is simulated for a fixed fault without any DG penetration. The table of results for this case is listed in table I. Nominal voltage =11kv voltage factor c=1.1(max) fault current is in Kilo Ampere

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I
SHORT CIRCUIT REPORT WITHOUT DG

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	1.934	1.944	1.944
Peak Current (KA,rms)	3.6	3.325	3.579	3.345
Breaking Current(KA,rms,Sym.)	1.96	1.934	1.944	1.944
Steady State Current (KA,rms)	1.96	1.934	1.944	1.944

B. Fault Analysis with DG

Wind turbine Generators having equal ratings of 2MW are considered as DG source for penetration at different nodes in the selected feeders. The four cases under studied are described as under.

Case 2: 2 MW DG located at Bus 10
The system is simulated for a fixed fault with 2MW DG source connected to a 400A load feeder bus comprising of 70Km length. The table of results for this case is listed in table II.

TABLE II
SHORT CIRCUIT REPORT WITH DG AT BUS 10

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	3.06	1.934	1.955	1.945
Peak Current (KA,rms)	5.91	3.326	3.588	3.345
Breaking Current(KA,rms,Sym.)	3.06	1.934	1.955	1.945
Steady State Current (KA,rms)	3.06	1.934	1.955	1.945

Case 3: 2 MW DG located at bus 23
The system is simulated for a fixed fault with 2MW DG source connected to 80A load feeder bus comprising of 50KM length. The table of results for this case is listed in table III.

TABLE III
SHORT CIRCUIT REPORT WITH DG AT BUS 23

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	3.222	1.944	1.956
Peak Current (KA,rms)	3.6	5.98	3.579	3.377
Breaking Current(KA,rms,Sym.)	1.96	3.222	1.944	1.956
Steady State Current (KA,rms)	1.96	3.222	1.944	1.956

Case 4: 2MW DG located at Bus 38
The system is simulated for a fixed fault with 2MW DG source connected to 80A, 70Km feeder. The result for this case is listed in table IV.

TABLE IV
SHORT CIRCUIT REPORT WITH DG AT BUS 38

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	1.934	2.099	1.945
Peak Current (KA,rms)	3.6	3.326	3.996	3.345
Breaking Current(KA,rms,Sym.)	1.96	1.934	2.099	1.945
Steady State Current (KA,rms)	1.96	1.934	2.099	1.945

Case 5:2 MW DG located at Bus 49
In this case 2MW DG source is connected to a feeder having 50km length and possessing 400A current. The results are listed in table 5

TABLE V
SHORT CIRCUIT REPORT WITH DG AT BUS 49

	L-G Fault Location			
	Bus 9	Bus 21	Bus 37	Bus 49
Initial Sym. Current (KA,rms)	1.96	1.947	1.944	3.722
Peak Current (KA,rms)	3.6	3.36	3.579	6.998
Breaking Current(KA,rms,Sym.)	1.96	1.947	1.944	3.722
Steady State Current (KA,rms)	1.96	1.947	1.944	3.722

V. Comparative Analysis and Discussion

This section describes the Comparison of initial symmetrical fault current, peak short circuit current and steady state current of the system during fault with and without the DG interconnection firstly while in the second section the comparison of fault currents between these feeders is discussed briefly.

A. Comparison of fault current with and without DG connected

The below chart shows the values of short circuit currents at the studied buses and is considered as base and set values for the protection settings of equipment used in the grid, those values will be compared with the values obtained from all other cases which contains a DG source.

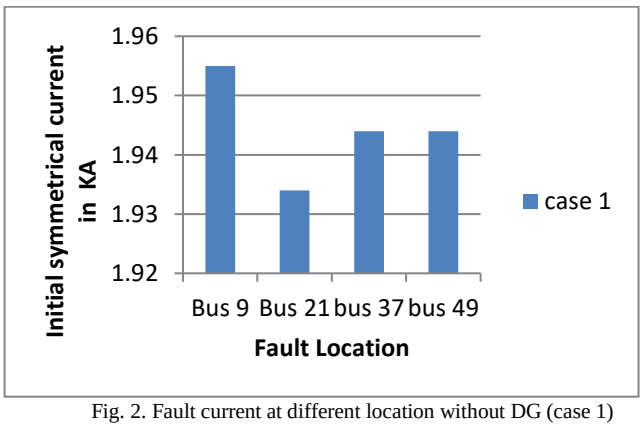


Fig. 2. Fault current at different location without DG (case 1)

To investigate the effect of DG on a feeder during fault conditions a series of four cases are considered which are compared with case 1; fig. 3 shows the comparison of case 2 with case 1.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

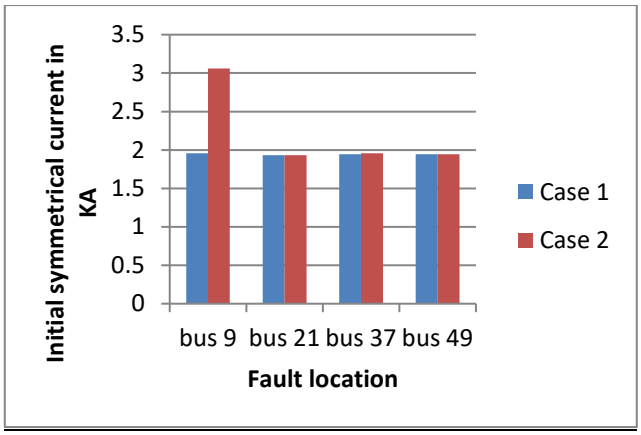


Fig. 3. Comparison between case 1 and 2

It shows that whenever DG is integrated at node 10 and a bolted fault occurs at bus 9 of the same feeder the magnitude of short circuit current increases from 1.955 KA to 3.061KA

In this case the parameters of feeder are different from the previous scenario, when a 2MW DG is brought into the system at node 23 and fault occurs at the bus 21 of the same feeder having length of 50KM and possessing 80A current, the results are compared with the base case as described in the fig. 4 below.

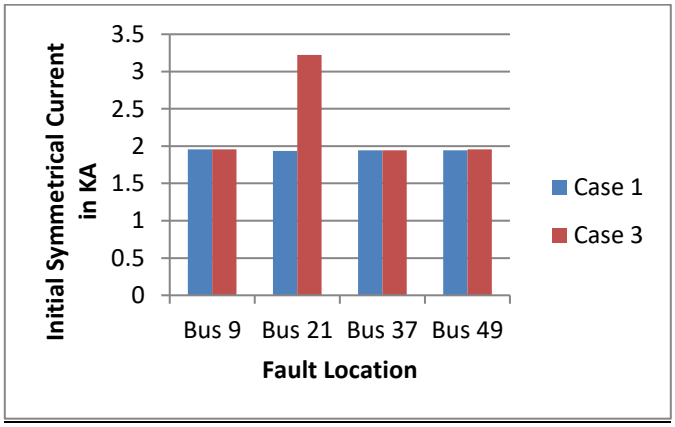


Fig. 4. Comparison between case 1 and 3

It is also evident from the comparison that addition of DG causes the short circuit contribution of feeder during fault conditions increased by collective 1.286KA.
The below comparison is validation of the objective that magnitude of short circuit current is increased during fault conditions with DG even when the length of feeder is maximum and minimum current of 80A is passing through it as shown in the fig. 5

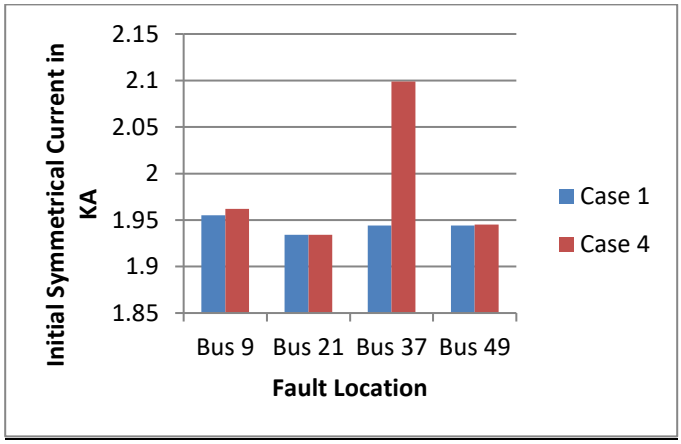


Fig. 5. Comparison between case 1 and 4

In the Last case DG is added at node 49 of a feeder comprising of 400A current and a total length of 70KM, the results in case 5 are compared with case 1 in the fig. 6 below.

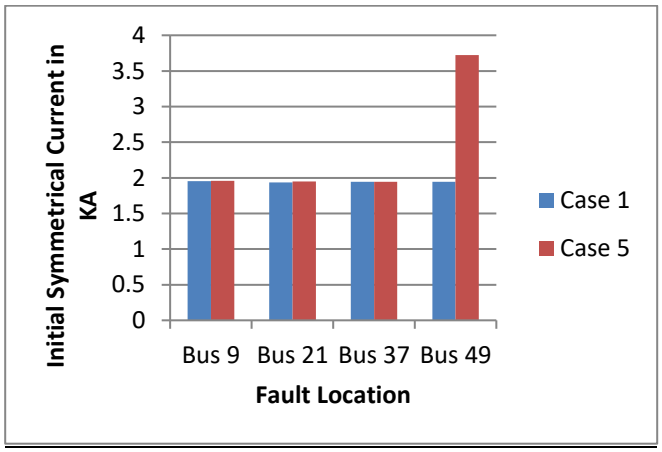


Fig. 6. Comparison between case 1 and 5

From the comparison it is extracted that whenever the location of fault and integrated DG is same the feeder contributes to maximum amount short circuit current during fault conditions, thus establishing a conclusion that whatever the parameter of a feeder may be an integrated DG will result in increase of short circuit current.

B. Comparison of Feeders containing DG

In this section two cases are considered for investigation to study the impact of DG on a feeder when its length is increased during fault conditions, the location of fault is maintained constant and short circuit currents are compared with each other as obtained in case 2 and case 5 which is described in the fig. 7

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

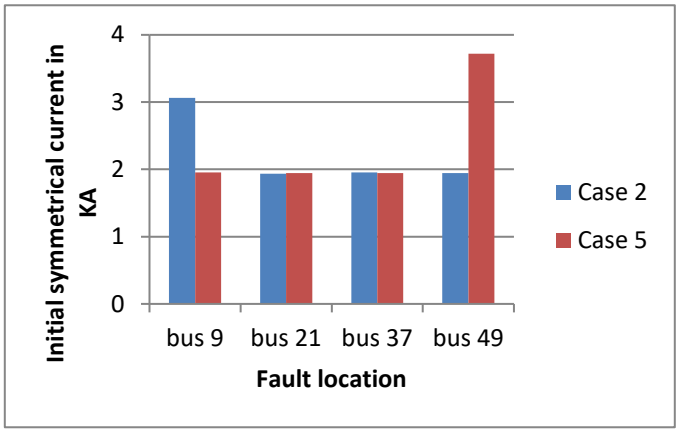


Fig. 7. Comparison between case 2 and 5

From the analysis it is obvious that when the length of a feeder is enhanced form 50km to 70Km while keeping maximum amperes of 400A flowing through it and fault occurs on the feeders integrated with DG, the short circuit current decrease from 3.722A to 3.061KA, so a decrease of 0.661KA suggests with conclusion that whenever the length of a line is increased irrespective of maximum or minimum load connected to it with DG integrated during fault, the fault current minimizes due to the fact that line impedance become larger with enhanced length thereby decreasing the effect of fault current on the distribution grid.

VI. Conclusion

Penetration of DG into a distribution system causes an increase in the fault level of the network at any fault location. In this paper Practical 132KV Grid Station has been considered as a case study for short circuit analysis with the DG connected. During the analysis it has been investigated that DG integration into the 11KV distribution feeder changes the initial symmetrical current contribution which can alter the protection configuration of the Grid station.

This paper also provides suitable location for DG integration in the selected feeders on the basis of its length suggesting that lengthy feeder has less short circuit current contribution and might be considered for penetration as it will not vary from the default protection setting of equipment's in the Grid.

VII. References

[1] Sarabia Angel Fernández. "Impact of distributed generation on distribution system." PhD dissertation Aalborg University, 2011.

[2] Jyotsna Sharma, Akhilesh Dobhal, " Analysis and Simulation of Faults in Power Distribution Systems with Distributed Generation *International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 9, September- 2014 ISSN: 2278-0181.*

[3] Renuka Kamdar, Mukesh Kumar Kirar, Manoj Kumar, Ganga Agnihotri, " Short Circuit Analysis of an Industrial Distribution System," *Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering* ISBN: 978-981-07-6260-5 doi:10.3850/ 978-981-07-6260-541.

[4] Hima A.Patel, Viabhav M.Sharma, and Anuradha Deshpande, " Relay Coordination using ETAP," *International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015 1583 ISSN 2229-5518.*

[5] Elmathana, Mohamed Talal Mohamed, Ahmed Zobaa, and George Smith. "The Effect of Distributed Generation on Power System Protection." (2010).

[6] Bong Eng Yueh, "Distributed Generation Impacts on the Coordination of Overcurrent Relay in Distributed Network," Universiti Teknologi Malaysia, Nov. 2012.

[7] "International standard IEC 60909, short circuit current in three phase a.c systems" International Electro technical commission, Jul. 2001.

[8] Khan, R. A. J., Junaid, M., & Asgher, M. M. (2009, November). Analyses and monitoring of 132 kV grid using ETAP software. In *Electrical and Electronics Engineering, 2009. ELECO 2009. International Conference on* (pp. I-113). IEEE.

[9] Keith Brown, Herminio Abcede, Farrokh Shokooh, "Interactive simulation of power system & ETAP application and Techniques" IEEE operation Technology, Irvine, California.

[10] Jillani, Syed Sagheer Hussain Shah. "Impact of distributed generation on distribution systems and its protection." (2012).

.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm

Abdeljalil EL ABDOULI, Larbi HASSOUNI, Houda ANOUN
RITM Laboratory, CED Engineering Sciences
Ecole Superieure de Technologie
Hassan II University of Casablanca, Morocco
elabdouli.abdeljalil@gmail.com, lhassouni@hotmail.com, houda.anoun@gmail.com

***Abstract*—Twitter is a web-based communication platform, which allows its subscribers to disseminate messages called "tweets" of up to 140 characters where they can share thoughts, post links or images. Therefore, Twitter is a rich source of data for opinion mining and sentiment analysis. The simplicity of use and the services offered by the Twitter platform allow it to be widely used in the Arab world and especially in Morocco, this popularity leads to an accumulation of a large amount of raw data that can contain a lot of valuable information. In this paper, we address the problem of sentiment analysis in Twitter platform. First, we try to classify the Moroccan users' tweets according to the sentiment expressed in them: positive or negative. Second, we discover the subjects related to each category to determine what they concern, and finally, we locate these "tweets" on Moroccan map according to their categories to know the areas where the tweets come from. To accomplish this, we adopt a new practical approach that applies sentiment analysis to Moroccan "tweets" using a combination of tools and methods which are: (1) Apache Hadoop framework (2) Natural Language Processing (NLP) techniques (3) Supervised Machine Learning algorithm "Naive Bayes" (4) Topic Modeling using LDA (5) Plotting tool for interactive maps called "Folium". The first task of our proposed approach is to automatically extract the tweets with emotion symbols (e.g., emoticons and emoji characters) because they directly express emotions regardless of used language, hence they have become a prevalent signal for sentiment analysis on multilingual tweets. Then, we store the extracted tweets according to their categories (positive or negative) in a distributed file system using HDFS (Hadoop Distributed File System) of Apache Hadoop framework. The second task is to preprocess these tweets and analyze them by using a distributed program written in Python language, using MapReduce of Hadoop framework, and Natural Language Processing (NLP) techniques. This preprocessing is fundamental to clean tweets from #hashtags, URLs, abbreviations, spelling mistakes, reduced syntactic structures, and many; it also allows us to deal with the diversity of Moroccan society, because users use a variety of languages and dialects, such as Standard Arabic, Moroccan Arabic called "Darija", Moroccan Amazigh dialect called "Tamazight", French, English and more. Afterward, we classify tweets obtained in the previous step using Naive Bayes algorithm into two categories (positive or negative), then we use the Topic Modeling algorithm LDA to discover general topics behind these classified tweets. Finally, we graphically plot classified tweets on our Moroccan map by using the coordinates extracted from them.**

Keywords: Apache Hadoop framework; HDFS; MapReduce; Python Language; Natural Language Processing; Supervised Machine Learning algorithm "Naive Bayes"; Topic Modeling algorithm LDA; Plotting tool for interactive maps.

I. INTRODUCTION

The emergence of Web 2.0 has led to an accumulation of valuable information and sentimental content in the Web; such content is often found in the comments of users of Social Network Platforms, in messages posted in discussion forums and product review sites, etc. The Twitter platform is very popular, and its users post a lot of comments to express their opinions, sentiments, and other information. This transforms twitter platform into a rich source of data for data mining and sentiment analysis. In this paper, we are interested in the sentiment analysis of the Moroccan users, we provide, below, some statistics on their activities. According to the Arab Social Media Report [1], which started in 2011 and aims to understand the impact of social media on societies, development, and governance in the Arab region, the monthly number of active users of the platform Twitter nearly doubled between 2014 and 2017. It went from 5.8 million to about 11.1 million. Regarding Morocco, the number of active users of the Twitter platform has grown of 146,300 users, in the last three years, to reach the number of 200 thousand users. Morocco thus ranks 9th among the Arab countries registering the highest number of users. These statistics prompted us to lead a study that aims to analyze the sentiments expressed in the tweets published by Moroccan users, despite the difficulties quoted before.

The primary aim of this research is to identify the sentiments contained in the tweets posted from the Moroccan region by proposing a new practical approach for analyzing the Moroccan user-generated data on Twitter. Our approach is based on a system, which automatically handles the streaming of the most recent tweets from Twitter platform using the open and accessible API of Twitter that returns well-structured tweets in JSON (JavaScript Object Notation) format. These tweets shape the training set, and are classified into two categories (Positive or Negative) according to the emotion symbols (e.g., emoticons and emoji characters) which exist in each tweet, then they are stored in our distributed system using HDFS [2]. These tweets are preprocessed by a distributed program using MapReduce [3], which is written in Python language using Natural Language Processing (NLP) techniques [4], and it's launched on MapReduce using the Pig UDF [5] (User Defined Functions). This preprocessing is fundamental to clean the tweets which are very noisy and contain all kind of spelling, grammatical errors and also to handle the linguistic diversity used by Moroccan users in the tweets. The result of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the previous step is a clean and filtered corpus of tweets that is divided into "Positive" (text with happy emoticons), and Negative" (text with sad and angry emoticons) samples. This corpus is used to form the training set for the Naive Bayes algorithm to identify the sentiment within the new collected tweets, then we apply topic modeling using LDA to discover the hidden topics within these tweets. Finally, we graphically plot the classified tweets using a tool called "Folium" on our Moroccan map by using the coordinates extracted from them, to discover the relationship between the areas of classified tweets and determined topics.

The remainder of the paper is organized as follows; we present some related work in Section II. In Section III, we introduce the tools and methods used to realize our system. In Section IV, we describe our system. Finally, in Section V; we end with a conclusion and work in perspective.

II. Related Work

Sentiment Analysis is receiving an increasingly growing interest from many researchers, which have begun to search various ways of automatically collecting training data and perform a sentiment analysis. [12] have relied on emoticons for defining their training data. [13] have used #hashtags for creating training data and they limit their experiments to sentiment/non-sentiment classification, rather than (positive-negative) classification. [14] have used emoticons such as ":-)" and ":-(" to form a training set for the sentiment classification, the author collected texts containing emoticons from Usenet newsgroups, and the dataset was divided into "positive" and "negative" samples. [15] have covered techniques and approaches that promise to enable the opinion-oriented information retrieval directly. [16] have used Twitter to collect training data and then to perform a sentiment search, they construct a corpus by using emoticons to obtain "positive" and "negative" samples and then use various classifiers, the best result was obtained by using Naïve Bayes Classifier.

III. Tools And Methods

A. Apache Hadoop

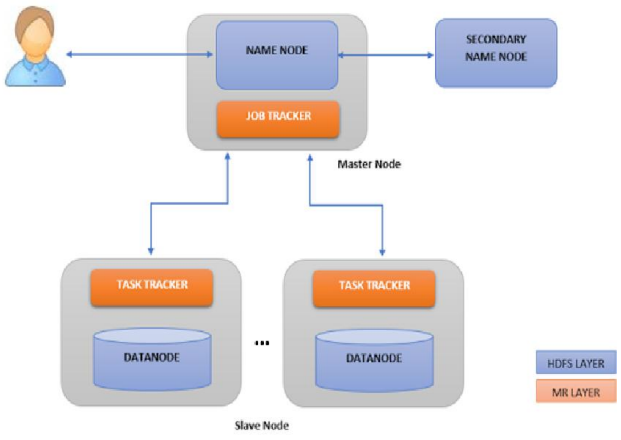


Figure 1. Apache Hadoop Architecture

Our approach is built using a specialized infrastructure, based on the Apache Hadoop Framework. The Apache Hadoop is an open-source software framework written in Java for processing, storing and analyzing large volumes of unstructured data on computer clusters built from commodity hardware.

The Hadoop Framework become a brand name, which contains two primary components. The first one is HDFS [5], which stands for Hadoop distributed file system; it is an open-source data storage, inspired by GFS (Google File System), it is a virtual file system that looks similar to any other file system, but the difference is that the file gets split into smaller files. The second one is MapReduce, which is an open-source programming model developed by Google Inc. Apache adopted the ideas of Google MapReduce and improved it. MapReduce provides a mechanism to break down every task into smaller tasks and the integration of results.

The HDFS (Hadoop Distributed File System) [2] system has many similarities with existing distributed file systems. However, the differences are significant, it is highly fault-tolerant and designed using low-cost hardware, also designed to be available and scalable. It provides high throughput access to stored data and can store massive files reaching the terabytes. By default, each stored file is divided into blocks of 64 MB, each block is replicated in three copies. The HDFS is based on Master and Slaves architecture in which the master is called the NameNode and slaves are called DataNodes, and it consists of:

a) Single NameNode: running as a daemon on the master node, it holds the metadata of HDFS by mapping data blocks to data nodes, and it is the responsible of managing the file system namespace operations.

b) Secondary NameNode: performs periodic checkpoints of the file system present in the NameNode and periodically joins the current NameNode image and the edits log files into a new image and uploads the new image back to the NameNode.

c) DataNodes: running as daemons on slave nodes, they manage the storing of blocks within the node (their default size is 128 MB). They perform all file system operations according to instructions received from the NameNode, and send a Heartbeat containing information about the total storage capacity of DataNode and Block report on every file and block they store to the NameNode.

The MapReduce [3] is the heart of Hadoop. It is a software framework that serves as the compute layer of Hadoop, it is modeled after Google's paper on MapReduce. It's characterized by fault tolerance, the simplicity of development, scalability, and automatic parallelization. It allows parallelizing the processing of massive stored data by decomposing the job submitted by the client into Map and Reduce tasks. The input of the Map task is a set of data as a key-value pair, and the output is another set of data as a key-value pair. The input of the reduce task is the output from a map task. Between the reduce input and the map output, MapReduce performs two essential operations, shuffle phase that covers the transformation of map outputs based on the output keys, and sort phase that includes the merge and sort of map outputs.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The MapReduce is also based on a master-slave architecture, and it consists of:

a) *JobTracker*: is running as a daemon on the master node, its primary role is accepting the job and assigning tasks to TaskTrackers running on slave nodes where the data is stored. If the TaskTracker fails to execute the task, the JobTracker assigns the task to another TaskTracker where the data are replicated.

b) *TaskTracker*: running as a daemon on slave nodes, it accepts tasks (Map, Reduce, and Shuffle) from JobTracker and executes program provided for processing. The TaskTrackers report the free slots within them to process data and also their status to the JobTracker by a heartbeat.

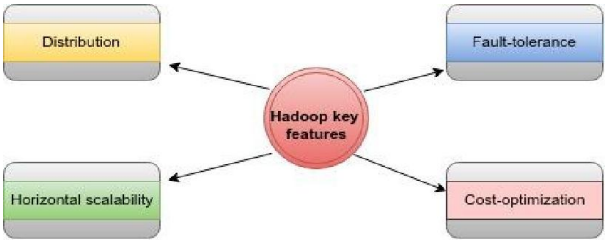


Figure 2. Key Features of Hadoop

The Key Features of Hadoop are:

Distribution: The storage and processing are spread across a cluster of smaller machines that work together.

Horizontal scalability: It is easy to extend a Hadoop cluster by adding new devices.

Fault-tolerance: Hadoop continues to operate even when a few hardware or software components fail to work correctly.

Cost-optimization: Hadoop runs on standard hardware; it does not require expensive servers.

Other Hadoop-related projects [7] at Apache that can be installed on top of or alongside Hadoop include:

- *Flume* [21]*:* is a framework for populating massive amounts of data into Hadoop.
- *Oozie* [22]*:* is a workflow processing system.
- *Mahout* [23]: Mahout is a data mining library.
- *Pig* [8]: a high-level data-flow language and execution framework for parallel computation.
- *Avro* [24]: a data serialization system.
- *HBase* [25]: a scalable and distributed database that supports structured data storage for large tables.
- *Hive* [26]: a data warehouse infrastructure that provides data summarization and ad hoc querying.
- *Spark* [27]*:* provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- And much more.

B. *Natural Language Processing (NLP)*

Natural Language Processing [4] is a part of computer science focused on developing systems that allow computers to recognize, understand, interpret and reproduce human language. NLP is considered as a subfield of artificial intelligence, and by using its algorithms, developers can perform tasks such as topic segmentation, translation, automatic summarization, named entity recognition, sentiment analysis, speech recognition, and much more.

There are two components of NLP. The first component is Natural Language Understanding (NLU) whose main function is to convert human language into representations that are easier for computer programs to manipulate. The other is Natural Language Generation (NLG) translate information from computer databases into readable human language. There are five steps in NLP:

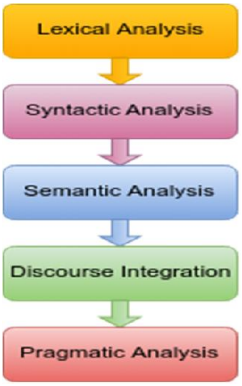


Figure 3. The steps of NLP

a) *Lexical Analysis:* identifying and analyzing the structure of words and dividing the whole text into paragraphs, sentences, and words.

b) *Syntactic Analysis:* analyzing and arranging words in a sentence in a structure that shows the relationship between them.

c) *Semantic Analysis:* extracting the exact meaning or the dictionary meaning of sentences from the text.

d) *Discourse Integration:* handles the meaning of current sentence depending on the sentence just before it.

e) *Pragmatic Analysis:* analyzing and extracting the meaning of the text in the context.

We use Natural Language Processing to perform tasks such as:

- Tokenization / segmentation
- Part of Speech (POS) Tagging: assign part-of-speech to each word.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Parsing: create a syntactic tree for a given sentence.
- Named entity recognition: recognize places, people...
- Translation: Translate a sentence into another language.
- Sentiment analysis.
- Etc.

Using the NLP is necessary for our system because tweets are characterized by a noisy text containing many unwanted data; in addition, the language diversity used in Moroccan society adds many difficulties to the processing of tweets' content generated by Moroccan users.

C. Scikit-learn and Naive Bayes algorithm

Scikit-learn [18] is an open source library for machine learning that is simple and efficient for data mining and data analysis for the Python programming language. It is Built on *NumPy, SciPy,* and *Matplotlib* [10]; it includes many algorithms for classification, regression and clustering algorithm, and more. Because it is a robust library, we choose to Implement naive Bayes classifier in python with *scikit-learn*.

The Naive Bayes [19] is a supervised classification algorithm based on Bayes' Theorem with an assumption that the features of a class are unrelated, hence the word naive. The Naive Bayes classifier calculates the probabilities for every factor; then it selects the outcome with the highest probability.

Preprocessed tweets with NLP is given as input to train input set using Naïve Bayes classifier, then, trained model is applied to new collected tweets to generate either positive or negative sentiment.

The Bayes theorem is as follows:

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

Where:

- P(H): the probability of the hypothesis H being true. This is known as the prior probability.
- P(E): the probability of the evidence (regardless of the hypothesis).
- P(E|H) is the probability of the evidence given that hypothesis is true.
- P(H|E) is the probability of the hypothesis given that the evidence is there.

There are many applications of Naive Bayes Algorithms:

- Text classification/ Spam Filtering/ Sentiment Analysis
- Recommendation Systems.
- Real-time Prediction: Naive Bayes is a fast classifier, and it can be used for making predictions in real time
- Multi-class Prediction: more than two classes to be predicted.

D. PIG UDF

Apache Pig [8] is a popular data flow language; it is at the top of Hadoop and allows creating complex jobs to process large volumes of data quickly and efficiently. It will consume any data type: Structured, semi-structured or unstructured. Pig provides the standard data operations (filters, joins, ordering).

Pig provides a high-level language known as Pig Latin for programmers who are not so good at Java. It is a SQL-like language, which allows developers to perform MapReduce tasks efficiently and to develop their functions for processing data.

A Pig UDF [5] (User Defined Functions) is a function that is accessible to Pig but written in a language that is not PigLatin like Python, Jython or other programming languages; it is a function with a decorator that specifies the output schema.

We use Pig UDF to execute NLP program, written with Python language in a distributed manner using MapReduce. In consequence, the preprocessing became very fast and spread over the stored tweets.

E. Topic Modeling Using Latent Dirichlet Allocation(LDA)

Topic modeling allows us to organize, understand and summarize large collections of textual information. It helps to discover hidden topical patterns that are present in the collection; annotate documents according to these topics; and use these annotations to organize, search and summarize texts.

Topic models are unsupervised machine learning algorithms, which allow discovering hidden thematic structure in a collection of documents. These algorithms help us to develop new ways of text exploration. Many techniques are used to obtain topic models, but the most used one is Latent Dirichlet Allocation (LDA) [17].

LDA algorithm works as a statistical machine learning and text data mining; it allows discovering the different topics in a collection of documents. It consists of a Bayesian inference model that calculates the probability distribution over topics in each document, where each topic is characterized by a probabilistic distribution based on a set of words.

The LDA algorithm is used in our system, to discover the topics of classified tweets (positive and negative). For this reason, we implement a free python library for LDA called "Gensim" [20].

F. Interactive maps using Folium

Folium [11] is a powerful Python library that allows visualizing geospatial data onto interactive maps; it provides the facilities to transform coordinates to different map projections. The visualization happens "inline" or within the Python environment, using *IPython Notebook* and the results are interactive which makes this library very useful for dashboard building.

The Plotting of classified tweets in Moroccan map is necessary to discover the general mood in Moroccan regions as well as the dominant topics by using LDA.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. Architecture Of The System

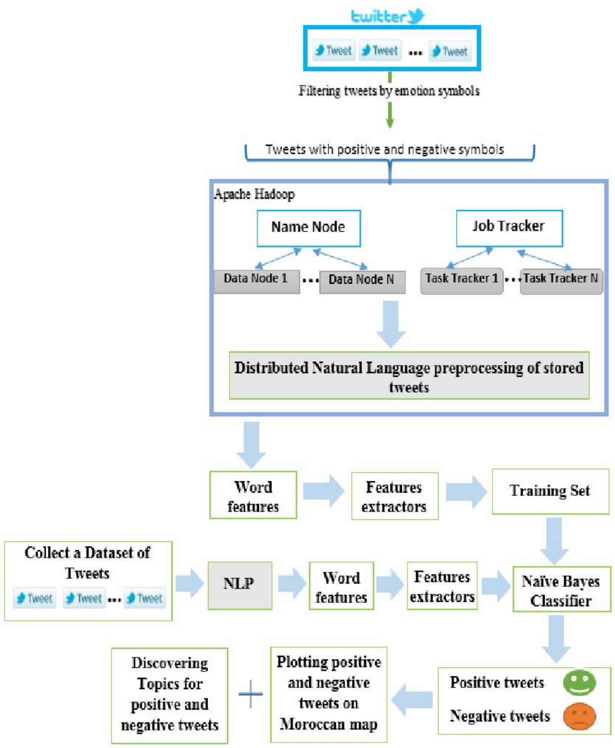


Figure 4. Architecture of the System

The first part of our system involves the extraction of data from the social network Twitter, which is an essential task in the data analysis process. All these tweets contain coordinates from different locations in Morocco and stored in the HDFS.

A. Data Description, Collection, and Storage

1) Data Description:

Twitter has unique conventions that make it distinct from other social network platforms; indeed tweets have many unique attributes, which differentiates Twitter from other platforms in the field of research. The First convention is the maximum length of a tweet, which is 140 characters, an average length of 14 words per tweet. The second convention is the availability of data using Twitter API; it is much easier to collect a large volume of tweets for training.

In Twitter messages, we can find acronyms, emoticons and other characters that express special meanings. The Emoticons can be represented using punctuation and letters or pictures; they express the user's mood. Emoticons can be categorized as:

- Happy emoticons (positive) :-) :) :D, etc
- Sad emoticons (Negative) :-(:(:-c, etc.

To use the method based on the emotion symbols we need to make an assumption. This assumption is that the emotion symbols in the tweet represent the overall sentiment contained in that tweet. This assumption is reasonable because in the majority of cases the emotion symbols will correctly represent the overall sentiment of that tweet as the maximum length of a tweet is 140 characters.

2) Data Collection:

Twitter Platform allows developers to collect data via Twitter API, but first, they need to create an account on *https://apps.twitter.com*. For each created account, Twitter provides four secret information: consumer key, consumer secret key, access token and access secret token, then we are authorized to access the database and retrieve tweets using the streaming API.

To get tweets that contain emoticons which are generated in Moroccan region, we filter tweets by location coordinates using the Streaming API and by a list of positive and negative emotion symbols. We get the geographical coordinates (latitude and longitude) of Morocco that we utilized in this filter, by using the specialized website in geolocation *http://boundingbox.klokantech.com*.

To handle the streaming of data from Twitter, we used Python library *Tweepy* [9] as shown in the script below that allows accessing to Twitter API.

```
import tweepy
import json
import hadoopy

consumer_key= XXX
consumer_secret= XXX
access_token= XXX
access_token_secret= XXX

hdfs_path_Pos = 'hdfs://master:54310/tweets_Positive/'
hdfs_path_Neg = 'hdfs://master:54310/tweets_Negative/'

POSITIVE = ["*O", "*-*", "*O*", "*o*", "* *",":P", ":D", ":d",
":p", ";P", ";D", ";d", ";p",":-)", ";-)", ":=)", ";=)", ":<)", ":>)",
";>)", ";=)", "=}", ":)", "(:;)","(;", ":}", "{:", ";}", "{;:]","[;",
":')", ";')", ":-3","{;", ":]",";-3", ":-x", ";-x", ":-X",";-X", ":-}", ";-
=}", ":-]",";-]", ":-.)","^_^", "^-^","😀", "😁","😃", "😄","😅","
😆","😇","😈","😉","😊","😋","😌","😍","😎","😏","😗","😘
","😙","😚","😛","😜","😝"]

NEGATIVE = ["":(", ";(", ":'(","=(", "={", "):", ");",")':", ")';", ")=",
"}=",";-{{", ";-{", ":-{{", ":-{",":-(", ";-(",":,)", ":'{","[:", ";]","😒
","😐","😑","😣","😞","😟","😠","😡","😔","😖","😢","😥","
😦","😧","😨","😩","😪","😫","😭","😰","😱"]

class StdOutListener(tweepy.StreamListener):
   def on_data(self, data):

      decoded = json.loads(data)
      tweet_txt = decoded["text"]

      if decoded['place']['country_code'] == 'MA':
           ...
```

```
        for tok in tweet_txt.split(" ")
            for emoji in POSITIVE:
                if emoji.decode('utf-8') in tok:
                hadoopy.put(localeFile ,hdfs_path_Pos)
                ...
stream.filter (locations=[-
17.2122302,21.3365321,0.9984289,36.0027875],async='true',enc
oding='utf8')
```

3) Data storage using HDFS

The storage of filtered tweets gathered from the Twitter API in HDFS is handled by using a Python wrapper for Hadoop called *Hadoopy* [6], which allows performing operations like reading and writing data from and to HDFS. We create Two folders in HDFS, one for the positive tweets (*'hdfs://master:54310/tweets_Positive/)'* and the other for the negative tweets (*'hdfs://master:54310/tweets_Negative/'*) as shown in the previous script.

B. Processing filtred tweets with NLP

A major issue which faces us when we are dealing with Twitter data is the informal style of the posts. Most tweets are written informally, contain many errors and abbreviations, and do not follow any grammatical rule. To minimize the effect of this informality on our classification, we will pre-process tweets, in order to clean them, before using them. We might find words misspelled, and therefore must be detected and corrected to evaluate sentiment more accurately.

Also, the linguistic diversity that characterizes the communication of Moroccan users on social network Twitter complicate the task of classification. To deal with this issue, we create a python file that contains a dictionary of words that we gathered manually, to transform words written in Moroccan dialect, or in a dialect of Berber Tamazight into Standard Arabic. These words could be written using the Arabic or French alphabet then we store it in each slave node of our cluster and imported inside the NLP script executed in these nodes. Below, a part of this file

```
#-*- coding: utf8 -*-
MoroccanDialects = [
("katbghi", u'تحب'),
("khas", u'يجب'),
("ban", u'يظهر'),
...
(u'إيييه', u'نعم'),
(u'عافاك' ,u'ارجوك'),
(u'خود' ,u'امسك'),
...
("zgizzi",u'يتجادل'),
("zigiz",u'يشغل'),
("werg ",u'يحلم ')]
```

Moroccan dialect written in French alphabet

Moroccan dialect written in Arabic alphabet

Moroccan Amazigh dialect "Tamazight"

The NLP step contains all the programs needed to preprocess the stored data, starting with parsing the tweets and extracting relevant information for our analysis, which are:

- *Text*: text of the tweet.
- *Lang*: language used by the user to write the tweet.
- *Coordinates:* location coordinates of a tweet.

The library used to preprocess tweets with NLP is the Natural language processing Toolkit (NLTK) [9], which is a set of open-source Python modules, allowing programs to work with the human language data. It involves capabilities for tokenizing, parsing, and identifying named entities as well as many more features; it also provides over 50 corpora and lexical resources such as WordNet and a set of text processing libraries.

We use the following steps for preprocessing the filtered tweets:

a) Delete unnecessary data: usernames, emails, hyperlinks, retweets, punctuation, possessives from a noun, duplicate characters, and special characters like smileys.

b) Shorten any elongated words (→ تكبير تكبيـــــــــــــــر)

c) Normalize whitespace (convert multiple consecutive whitespace characters into one whitespace character).

d) Convert hashtags into separate words, for example; the hashtag #SentimentAnalysis is converted into two words Sentiment and analysis.

e) Create a function to detect the language used to write the text of tweet (Standard Arab, French or English).

f) Create a function for automatic correction of spelling mistakes.

g) Create a list of contractions to normalize and expand words like (*What's=>What is)*

h) Delete the suffix of a word until we find the root. For example (*Stemming => stem)*

i) Remove tokens of part of speech that are not important to our analysis by using the Part-Of-Speech software of Stanford University. This software reads the text and assigns parts of speech (noun, verb, adjective) to each word.

j) Remove stopwords of standard Arabic *(أن, إن, يعد...)*, French (*alors, à, ainsi, ...*), and English *(about, above, almost, ...)*.

These steps are assembled in a python file called *NLTK_Tweet.py*. This file is executed in a distributed manner by an Apache Pig file called *Pig_Tweet.pig*. The file *NLTK_Tweet.py* needs to be registered in the script of the Pig file using *Streaming_python* as follows:

```
REGISTER 'hdfs://master:54310/apps/NLTK_Tweet.py' USING
streaming_python AS nltk_udfs;
```

The launch of our file *NLTK_tweet.py* is defined as follows:

```
data = LOAD '/tweets_Positive /*  using TextLoader() AS
(line:chararray);

Result =  FOREACH data GENERATE
nltk_udfs.NLTK_Function(line));
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

C. Naïve Bayes Classifier

1) Data

Using Twitter API, we were able to collect experimentally a sample of 700 tweets (divided into positive and negative tweets) based on the emotion symbols and location filter, and 230 tweets as test set for accuracy evaluation of our classifier. All collected tweets are stored in a distributed manner using HDFS. The purpose of this paper, among others, is to be able to automatically classify a tweet as a positive or negative tweet. The classifier needs to be trained, that is why we use the stored tweets as training set after preprocessing step with NLP.

2) Implementation

For example, a fragment of the list of positive tweets looks like:

```
pos_tweets = [('I love this song, 'positive'),
              ('This picture is wonderful, 'positive'),
              ('I feel great this evening, 'positive'),
              ('This is my favorite food', 'positive')]
```

A fragment of the list of negative tweets looks like:

```
neg_tweets = [('I do not like this song, 'negative'),
              ('This picture is horrible', 'negative'),
              ('I feel sad this evening, 'negative'),
              (I hate this food, 'negative')]
```

We take these two lists and create a single list of tuples each containing two elements. The first element is an array containing the words and the second element is the type of sentiment. We ignore the words smaller than two characters, and we use lowercase for everything. The code is as follows:

```
tweets = []
for (words, sentiment) in pos_tweets + neg_tweets:
    words_filtered = [e.lower() for e in words.split() if len(e) >= 3]
    tweets.append((words_filtered, sentiment))
```

The tweets list now looks like this:

```
tweets = [
    (['love', 'this', 'song''], 'positive'),
    (['this', 'picture', 'wonderful'], 'positive'),
    (['feel', 'great', 'this', evening], 'positive'),
    (['this', 'favorite', 'food'], 'positive'),
    (['not', 'like', 'this', 'song'], 'negative'),
    (['this', picture, 'horrible'], 'negative'),
    (['feel', sad, 'this', evening], 'negative'),
   ([hate, 'this', 'food'], 'negative')])
]
```

3) Classifier

The list of word features needs to be extracted from the tweets. It is a list of every distinct word ordered by the frequency of occurrences. We use the following function and the two helper functions to get the list.

```
word_features = get_word_features(get_words_in_tweets(tweets))

def get_words_in_tweets(tweets):
    all_words = []
    for (words, sentiment) in tweets:
      all_words.extend(words)
    return all_words

def get_word_features(wordlist):
    wordlist = nltk.FreqDist(wordlist)
    word_features = wordlist.keys()
    return word_features
```

If we take a pick inside the function *get_word_features*, the variable 'wordlist' contains:

```
<FreqDist:
'this': 7,
'song': 2,
'feel': 2,
'evening': 2,
'picture': 2,
'wonderful': 1,
'favorite': 1,
'food':1
...
>
```

The list of word features is as follows:

```
word_features = [
'this',
'song',
'feel',
'evening',,
'picture',
'wonderful',
'favorite',
'food':1
    ...
]
```

The results show that 'this' is the most used word in our tweets, followed by 'song, then 'fell and so on …

We need to choose what features are pertinent to create our classifier. First, we need a feature extractor that returns a dictionary of words that are contained in the input passed. In our case, the input is the tweet. We use the word features list defined above along with the input to create the dictionary.

```
def extract_features(document):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains(%s)' % word] = (word in
document_words)
    return features
```

For example, let's call the feature extractor with the first positive tweet ['love', 'this', 'song']. We obtain the following dictionary which indicates that the document contains the words: 'love', 'this' and 'song'.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
{'contains(love)': True,
 'contains(evening)': False,
 'contains(this)': True,
 'contains(picture)': False,
 'contains(wonderful)': False,
 'contains(song)': True,
 'contains(favorite)': False,
 'contains(food)': False,
 'contains(horrible)': True,
 'contains(hate)': False,
 'contains(sad)': False,}
```

We use the method *apply_features* to apply the features to our classifier, and we pass the list of tweets along with the feature extractor defined above.

```
training_set = nltk.classify.apply_features(extract_features,
tweets)
```

The variable called 'training_set' contains the labeled feature sets, it is a list of tuples, where each tuple containing the feature dictionary and the sentiment category for each tweet.

```
[({'contains(love)': True,
   ...
   'contains(this)': True,
   ...
   'contains(song)': True,
   ...
   'contains(hate)': False,
   'contains(sad)': False},
  'positive'),
 ({'contains(love)': False,
   'contains(picture)': True,
   ...
   'contains(this)': True,
   ...
   'contains(wonderful)': True,
   ...
   'contains(hate)': False,
   'contains(sad)': False},
  'positive'),
 ...]
```

Now we can train our classifier using the training set.

```
classifier = nltk.NaiveBayesClassifier.train(training_set)
```

4) Testing the Classifier

To check the quality of our classifier by using the test set, we use an accuracy method in nltk that computes the accuracy rate of our model. Our approach reaches an accuracy of 69% which is considerate as a good value in our case.The simplest way to improving the accuracy of our classifier would be to increase the size of the training set.

```
import nltk.classify.util
print 'accuracy:', nltk.classify.util.accuracy(classifier, testTweets)
```

5) Classification of new collected tweets

Now that we have our classifier initialized and ready, we can try to classify collected and preprocessed tweets using NLTK and see what is the sentiment category output (positive or negative). Our classifier can detect that tweets have positive or negative sentiments. We evaluate our approach by streaming new collected tweets from Twitter API estimated at 300 tweets. A sample of collected tweets is as follows:

الوداد نسي ماتش الأهلي و بدأ يفكر في ماتش باتشوكا المكسيكي في كأس العالم و المصريين لازالو يحللون الهدف هل شرعي أو... https://t.co/PibiSBFoms from @monsef_filali at 11/07/2017 14:33

@lescitoyensorg Et ça continue ... from @cramounim at 11/07/2017 19:59

أداء مؤثر جدا و في قمة الخشوع لشيخي ماهر رعاه الله و حفظه مليكة @YouTube from @khadimarrahmane at 11/07/2017 15:39

Watching winner slowly realise that they're being kidnapped is the funniest thing ever #WinnerOverFlowers from @winneroediya at 11/07/2017 15:29

…

The below code is used to classify these new collected tweets using the classifier.

```
import nltk
from nltk.probability import FreqDist, ELEProbDist
from nltk.classify.util import apply_features,accuracy
...

print classifier.classify(extract_features(tweet.split()))
```

The output of the classification is the sentiment category of each tweet which is positive or negative. Our approach show good result despite the difficulties of multilingual tweets. some tweets are misclassified but we can override this issue by increasing the number of tweets in training set.

D. Topic Modeling with LDA

LDA is a probabilistic model used to determine the covered topics using the word frequency in the text. We use LDA in our approach for the classified tweets for each category(positive and negative). The LDA step will explain the reasons for the Moroccan user's mood. To generate the LDA model, we need to construct a document-term matrix with a package called "Gensim", which allows us to determine the number of occurrences of each word in each sentiment category. The LDA program used to discover topics is as follows:

```
from gensim import corpora, models
import hadoopy

fname_in = '/home /corpusTweetsSeniment.csv'
documentsPos = ""
documentsNeg = ""

with open(fname_in, 'rb') as fin:
    reader = csv.reader(fin)
    for row in reader:
        if row[3] == "positive":
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
        documentsPos = documentsPos + row[2] + ","
    elif row[3] == "negative":
        documentsNeg = documentsNeg + row[2] + ","

documentsPos =  documentsPos[:-1]
documentsNeg =  documentsNeg[:-1]

print ( ---- Topics  for positive tweets -----)
texts = [word.split() for word in documentsPos.split(",")]
dictionary = corpora.Dictionary(texts)
corpus = [dictionary.doc2bow(text) for text in texts]
lda = models.ldamodel.LdaModel(corpus, id2word=dictionary,
num_topics=2 , passes=10)
lda.show_topics()

print ( ---- Topics  for negative tweets -----)
texts = [word.split() for word in documentsNeg.split(",")]
dictionary = corpora.Dictionary(texts)
corpus = [dictionary.doc2bow(text) for text in texts]
lda = models.ldamodel.LdaModel(corpus, id2word=dictionary,
num_topics=2 , passes=10)
lda.show_topics())
```

For instance, the topics detected by our LDA model are:

```
---- Topics  for positive tweets ------
Topic #1: Maroc, football, equipe, russie, qualification
Topic #2: كأس، افريقيا، الوداد، جماهير

---- Topics  for negative tweets -----
Topic #1: تلوث، تهديد، سكان، بيئة، إشكالية
Topic #2: accident, circulation, mort, blesses, route
```

E. Plotting the classified tweets on map using Folium

During the streaming of filtered tweets from the Twitter API, we extract the coordinates (longitude and latitude) of each tweet. We then use these coordinates in Folium to show locations of tweets on our Moroccan map. The tweets that belong to positive mood are in green color and the negative mood are in red color. The developed program is as follows:

```
import folium
import csv

filename = '/home /corpusTweetsCoordinatesSeniment.csv'

map = folium.Map(location=[36.0027875,-17.2122302],
zoom_start=6)

  with open(filename) as f:
      reader = csv.reader(f)
      for row in reader:
        if row[3] == "positive":
          folium.Marker(location=[row[1],row[0]],popup=row[
          2],icon = folium.Icon(color ='green')).add_to(map)

        elif row[3] == "negative":
          folium.Marker(location=[row[1],row[0]],popup=row[
          2],icon = folium.Icon(color ='red')).add_to(map)

map.save("/home/abdeljalil/map_tweets.html")
map
```

The Figure 4 below shows the result of plotting classified tweets on the Moroccan map:

Figure 5. Locations of classified tweets on Moroccan map

This representation gives an idea about the locations of the Moroccan positive and negative tweets. This map and the topics generated by LDA are a good and perfect combination to study the mood of the Moroccan users, and more specifically to answer the two questions : Why this mood (LDA) and Where (Map).

V. Conclusion and Future Work

Twitter nowadays became one of the major tools and new types of the communication. People directly share their opinions through Twitter to the public. One of the very common analyses, which can be performed on a large number of tweets, is sentiment analysis. In the proposed work, we have presented a method for an automatic collection of a corpus that can be used to train a multilingual sentiment classifier so that it will be able to classify tweets into positive and negative. This classification is based on Naive Bayes classifier. Then we use methods to get insight from the classified tweets as the hidden topics and the locations of positive and negative tweets, which can conduct to better understanding of the Moroccan mood about different subjet and events. As future work, we plan to increase the accuracy of our classifier by increasing the number of filtered tweets, and by improving the preprocessing with NLP.

VI. References

[1] arabsocialmediareport, "Twitter in Arab Region". [Online]. Available: http://www.arabsocialmediareport.com/Twitter/LineChart.aspx. [Accessed: 01- Sep- 2017].

[2] Mrudula Varade and Vimla Jethani, "Distributed Metadata Management Scheme in HDFS", International Journal of Scientific and Research Publications,Volume 3, Issue 5, May 2013.

[3] M. Ghazi and D. Gangodkar, "Hadoop, MapReduce and HDFS: A Developers Perspective", Procedia Computer Science, vol. 48, 2015.

[4] M. Nagao, "Natural Language Processing and Knowledge", 2005 International Conference on Natural Language Processing and Knowledge Engineering.

[5] Pig.apache.org, "User Defined Functions". [Online]. Available: https://pig.apache.org/docs/r0.9.1/udf.html. [Accessed: 06- Sep -2017].

[6] "hadoopy", hadoopy.readthedocs.org. [Online]. Available: https://hadoopy.readthedocs.org/en/latest/. [Accessed: 01- Sep- 2017].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[7] "Apache Hadoop", [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hadoop.html. [Accessed: 01- Mar- 2017]

[8] pig.apache.org, "Welcome To Apache Pig". [Online]. Available: https://pig.apache.org/. [Accessed: 01- Sep- 2017].

[9] nltk.org/, "nltk". [Online]. Available: http://www.nltk.org/. [Accessed: 01- Sep- 2017].

[10] matplotlib.org/basemap/, " Welcome to the Matplotlib Basemap Toolkit documentation". [Online]. Available: http://matplotlib.org/basemap/. [Accessed: 01- Sep- 2017].

[11] "Folium", [Online]. Available: https://github.com/python-visualization/folium. [Accessed: 01- Mar- 2017].

[12] A. Pak, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proc. of LREC, 2010.

[13] D. Davidov, O. Tsur and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys", Proceedings of Coling, 2010

[14] Jonathon Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification", In ACL. The Association for Computer Linguistics, 2005.

[15] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval, 2008.

[16] Alec Go, Lei Huang, and Richa Bhayani, "Twitter sentiment analysis", Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group, 2009

[17] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research, 2003.

[18] scikit-learn.org/stable/, "scikit-learn". [Online]. Available: http://scikit-learn.org/stable/. [Accessed: 01- Sep- 2017].

[19] Z. Zhang, "Naïve Bayes classification in R", Annals of Translational Medicine, vol. 4, no. 12, pp. 241-241, 2016.

[20] radimrehurek.com/gensim/, "Gensim topic modeling for humans". [Online]. Available: https://radimrehurek.com/gensim/. [Accessed: 01- Sep- 2017].

[21] flume.apache.org/, "Apache Flume". [Online]. Available: https://flume.apache.org/. [Accessed: 01- Sep- 2017].

[22] oozie.apache.org/, " Apache Oozie Workflow Scheduler for Hadoop". [Online]. Available: http://oozie.apache.org/. [Accessed: 01- Sep- 2017].

[23] mahout.apache.org/, "Apache Mahout: Scalable machine learning and data mining". [Online]. Available: http://mahout.apache.org/. [Accessed: 01- Sep- 2017].

[24] avro.apache.org/, "Welcome to Apache Avro!". [Online]. Available: https://avro.apache.org/. [Accessed: 01- Sep- 2017].

[25] hbase.apache.org/, "Apache HBase – Apache HBase™ Home". [Online]. Available: https://hbase.apache.org/. [Accessed: 01- Sep- 2017].

[26] hive.apache.org/, "APACHE HIVE TM". [Online]. Available: https://hive.apache.org/. [Accessed: 01- Sep- 2017].

[27] spark.apache.org/, " Apache Spark™ - Lightning-Fast Cluster Computing". [Online]. Available: https://spark.apache.org/. [Accessed: 01- Sep- 2017].

Using biometric watermarking for video file protection based on chaotic principle

Saja J. Mohammed
Computer Science Department
College of Computer Science & Mathematics /Mosul University
Mosul/ IRAQ
Sj_alkado@yahoo.com

Abstract- **In the present time, security in the content of multimedia became one of significant science types. Watermarking is one type of multimedia protection, it is idea of protect digital components. Watermarking has extended and applied for many requirements, like fingerprinting, copyright protection, content indexing and many others watermarking application.**

The suggested algorithm is to hide a bio-watermarking encrypted data using video file as a cover in order to achieve video file protection. The recipient will need only to follow the required steps to retrieve the data of watermark. The idea of proposed method is based on hiding the watermark in audio partition of video file instead of video's image. Also use multiple frequency domains to hide the biometric watermark data using chaotic stream as key for encrypting the watermark and choose location for hiding. Subjective and objective tests (SNR, PSNR and MSE) are used to estimate the performance of the suggested method with applying simple attack that may attack the cover file.

Experimental result of the algorithm shows good recovering of watermark code which is virtually undetectable within the video file.

Keywords: video watermarking, DCT, DWT, Biometric system, chaotic.

I. INTRODUCTION

Nowadays, the digital media and the Internet have become so popular. That led to rise the requirements of secure data transmission. A number of useful techniques are proposed and already in use [1]. Watermark is one of these techniques which is a digital code embedded into the content of digital cover i.e. text, image , audio or video sequence [2].

Watermarking method is describe in the process as follows: Firstly, the abstraction of copyright data in the form of watermarks and imbedded in multimedia carriers using one of many embedding algorithms. After that, these carriers are distributed by the network or any digital storage. When necessary , the carriers are processed to detect the watermark existence . It is also possible to extract watermark for many various purposes[3].

In general, watermarking process is to embed some copyright data into the host data as an evidence ownership right. It must meet requirements which is: Security Obviously, Robustness, Imperceptibility and Capacity [4].

Various algorithms of digital video watermarking have been suggested. These techniques are categorized according to the domain which they working with. Some of these techniques embedded the watermark using the spatial domain by modification of the pixel values in each extracted video frame. These methods are entrusted to attacks and signal distortions. However, other techniques using the frequency domain to embed their watermark, this is the better robust to distortions[2].

Digital video is a sequence of still images merging with audio. The watermark will carry all types of information however the quantity of watermark data is limited. The vulnerability of the data is direct concerning of the amount of the information that carried by the watermark. The amount is absolutely limited by the size of particular video sequence[2].

II. WHAT IS BIOMETRICS?

Biometrics, is the process of authentication which depend on the physiological or behavioral properties and its ability to identify whether the person is authorized or not. Biometric properties distinctive as they can not be lost or forgotten, the presentation of identifying person will be done physically [5][6].

There are many of biometrics like fingerprint, face, hand thermogram, , signature, retina, iris, hand geometry, voice and so… .The most proven method is Iris -based identification . Iris can be defined as the colored part of eye, Fig. 1 shows the iris contents .The two eyes iris of any person have various iris pattern. Because the iris has a lot of characteristic which help to distinguish one iris from another, two conformable twins also have various iris patterns. Iris stills in a stable pattern not depended to the age affection that mean it stay in stability from the birth to the death. Also, the system of iris recognition can be un-invasive to their user[5][7].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

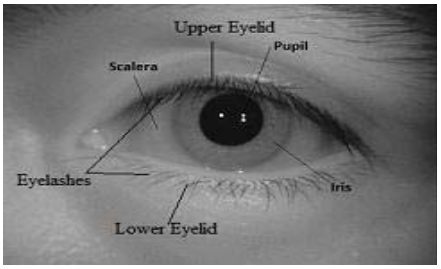


Figure 1. Structure of Iris

III. CHAOTIC SIGNAL

The chaotic signal is similar to noise signal, but it is certain in complete, that means if anyone has the initial values and the used function, that will be reproduce the same amount exactly. The profit of chaotic signal are[8]:

I. The initial conditions sensitivity

A minor variation in initial amount will cause important distinction in subsequent measures. The final signal will be differ completely if there is a small modification in the signal amount.

II. The accidental feature apparently

To compare with productive casual natural number in which the numbers scope cannot be generated again, the technique used for generating the same casual number in methods based on the chaotic function will create the ground that if the initial values and the used function are the same, the same number generated again.

III. The work deterministic

However, the chaotic functions were the casual manifest, they are wholly similar. That is if the initial values and the used function are fixed, the amounts of numbers will generate and re generate which seemingly have not any order and system. Logistic Map signal is one of the farthest known chaotic signals, this signal is presented by equation shown in (1):

$$X_{n+1} = rX_n (B\text{-}x_n) \qquad (1)$$

Where Xn gets the numbers in range [0,1]. The signal explain three various chaotic characteristics in three various ranges on the division of r parameter, the signal characteristics will be the best by assuming X0 =0.3.

- in $r \in [0,3]$, the signal characteristics in the first 10 iteration show some chaos and they were fixed after that , Fig. 2 (a)[9][10]
- in $r \in [3, 3.57]$, the signal characteristics in the first 20 iteration show some chaos , they were fixed after that, Fig. 2(b),
- in $r \in [3.57,4]$, the signal characteristics are chaotic in complete, Fig. 2(c)

Agreement with the above description and the requirements of the proposed algorithm to ensure complete chaotic characteristics for video watermarking, the logistic map chaotic signal with primary value X0=0.3 and $r \in [3.57,4]$ are used[9].

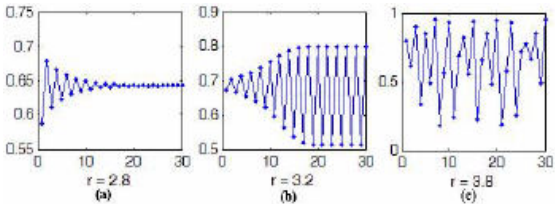


Figure. 2 The signal of logistic map chaotic with X0=0.3 and

$r \in [0,3]$, (b) $r \in [3, 3.57]$, (c) $r \in [3.57,4]$

IV. THE RELATED WORKS

There are many of watermarking methods based on video file as cover suggested in last period . One of these methods was proposed by Mobasseri (2000), who suggest a watermarking algorithm in compressed videos using spatial domain. Where Hong et al (2001) proposed an algorithm based on DWT they modify middle frequencies in the file. In other side Liu et al (2002) suggested a video watermarking algorithm using DWT to embed multi information bits. Chang & Tsai (2004) suggested a watermarking algorithm for a compressed video by VLC decoding and VLC code substitution. Zhong & Huang (2006) suggested video watermarking schema using spread-spectrum method for watermarking robustness improvement. Mirza et al (2007) suggest a video watermarking method using Principal Component Analysis [4].

V. THE PROPOSED METHOD

As we know video file format contain major two part of multimedia types: image and audio. It is generated by mixing the two kinds of multimedia types. The proposed method differs from the typical watermarking scheme. It is based on hiding watermark data in video's audio part instead of image one.

There are two categories of Digital watermarking technique: spatial domain watermarking technique and frequency domain watermarking techniques. The spatial domain methods hide the watermark using modifying some values of video file in directly way . The frequency domain technique will be embedding the watermark in best ways to ensure better determine of perception criterion and robust watermarking. Therefore the proposed algorithm used frequency domain to hide watermark data and in order to achieve more security multiple type of frequency domains with chaotic key are used.

In the proposed method, the watermark is based on biometrics (exactly on iris) to generate the watermarking code. The following sections discuss the proposed video Watermarking in details.

A) The proposed algorithm of embedding watermark code:

The proposed algorithm can be divided into two basic parts: generating the biometric watermark code and hiding it in video file data using chaotic key.

- *Generating the biometric watermarking code:*

To generate iris watermark data the iris (included in eye image) must be segmented .This will be made in the following steps : edge detection, circle detection and eyelid

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

detection. There are many technique for edge detection. This paper used canny edge detection and Hough transform to find iris and pupil boundaries. Iris image must be available in sender and receiver sides. For more security the watermark is encrypted using chaotic key.

The proposed algorithm of generating the bio-watermarking code is explained in the following steps:

Input: Iris image.

Output: Encrypted bio-watermarking code.

1) Begin
2) Choose iris image.
3) Apply iris segmentation.
4) Take iris data which is laying under pupil circle.
5) Apply edge detection using canny filter.
6) Generate chaotic key.
7) Encrypt iris data using the generated chaotic key.
8) End.

Fig. 3 shows the flowcharts of generating the bio-watermark code.

- *Embedding the watermark in video file using chaotic key:*

Input: Video file, Bio-watermark code.

Output: Watermarked video file.

1) Begin.
2) Choose video file to be cover file.
3) Split image and audio in it and consider audio part as a cover.

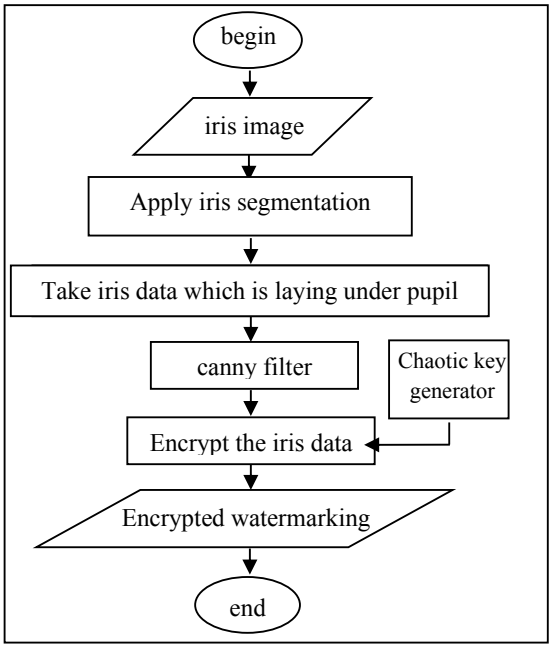


Figure 3. Generating the bio-watermarking code.

4) Apply DWT on audio part.
5) Apply DCT on resulted DWT coefficients.
6) Hide the length of watermark (Len) in first 4 bytes of cover data.
7) Generate chaotic key to be the index of chosen cover data .
8) Hide watermark code in cover by exchanging the fourth decimal number after comma in cover by another digit of watermark code.
9) Repeat this step until last digit in watermark code.
10) Apply DCT inverse, then DWT inverse.
11) Reformat the video cover.
12) End

Fig. 4 shows the proposed algorithm of hiding the biometric watermarking code in video file using chaotic key.

B) The proposed algorithm of extracting watermark code:

Input: The covered video file.

Output: Achieve video file protection or not.

1) Begin.
2) Input the covered video file.
3) Extract audio part from the covered video file.
4) Apply DWT on audio part.
5) Apply DCT on resulted DWT coefficients
6) Extract the length (Len) of watermark from first 4 byte in cover.
7) Generate chaotic key(for extracting and decryption operation).
8) Using the chaotic key to extract watermark code.
9) Repeat this step until reaching the length of watermark code.
10) Decrypt the extracted watermark using same chaotic key.
11) Independently… Generate the iris watermark code (origin one) by executing the steps of generating the biometric watermark (1 to 5).
12) Use the coparition between the onigin watermark with the extracted watermark data. If they are identical ,video file protection is achieved otherwise the file is not protected.
13) End

Fig.5 shows the proposed algorithm of extracting watermark code.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

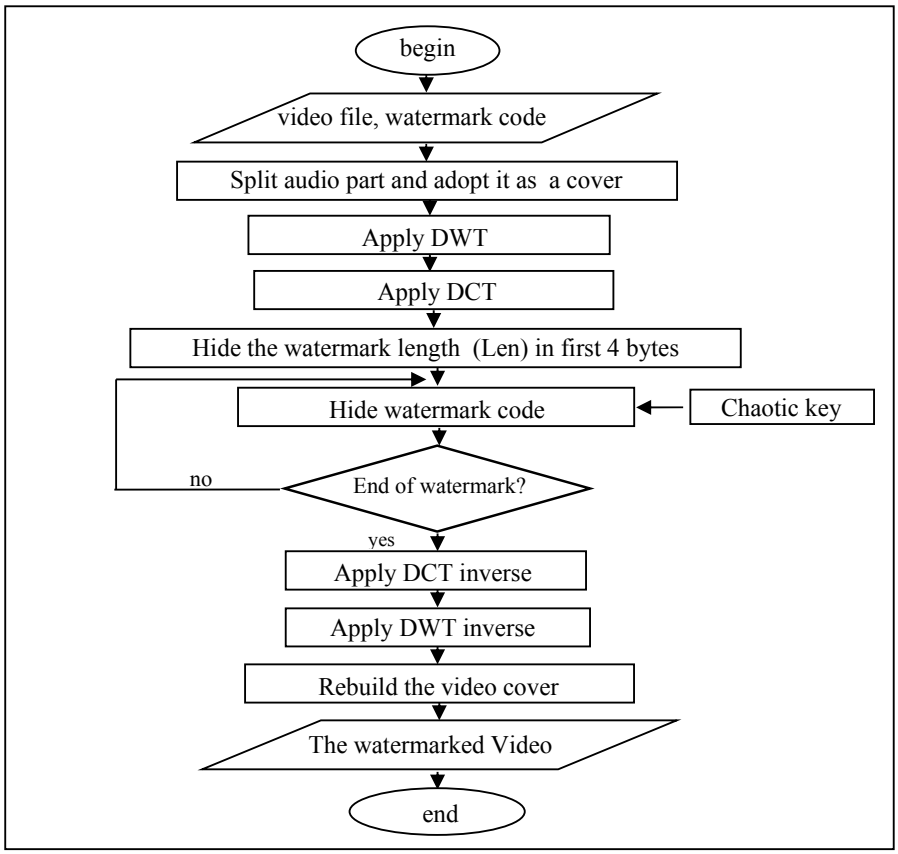


Figure. 4 The proposed algorithm of Hide the watermark in video file using chaotic key

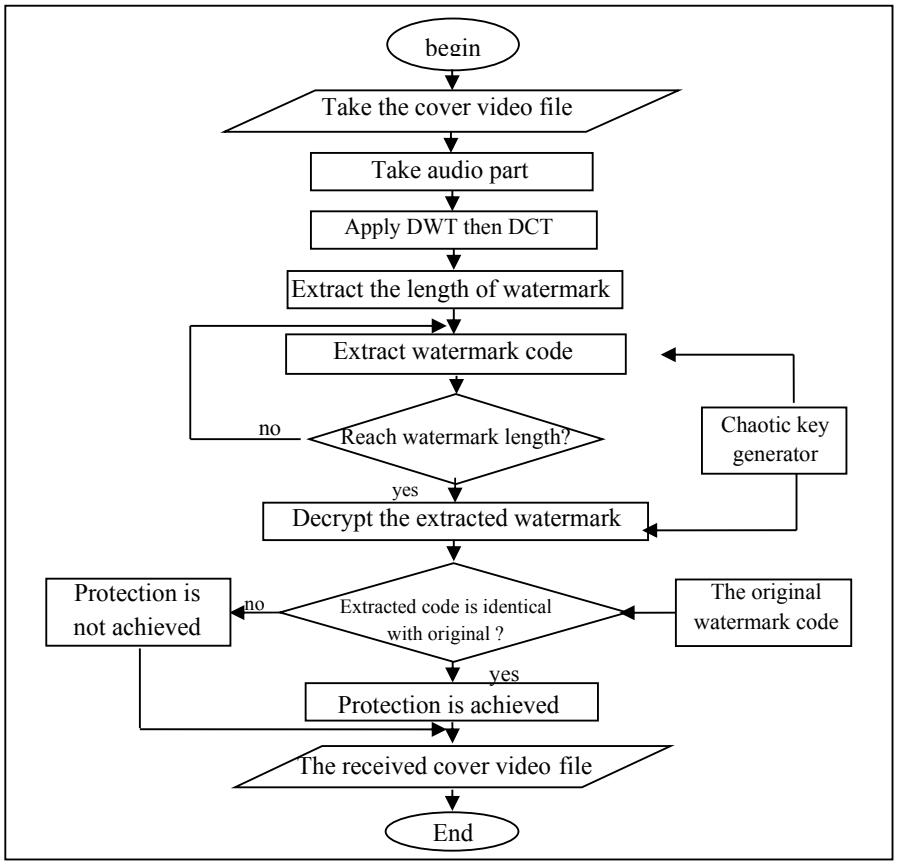


Figure 5. The proposed algorithm of extracting watermark code

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI. EXPERIMENTAL APPLICATION AND RESULTS

A number of video sequences have been tested using the proposed method. The bio-watermark is extracted from the watermarked video and its robustness is checked by calculating some famous measures.

Moreover, the proposed method is applied on many iris images obtained from CASIA database. At last the iris code is obtained and hidden in video file. Figs 6,7,8 show the experimental steps that are done on iris image to get bio-watermark code.

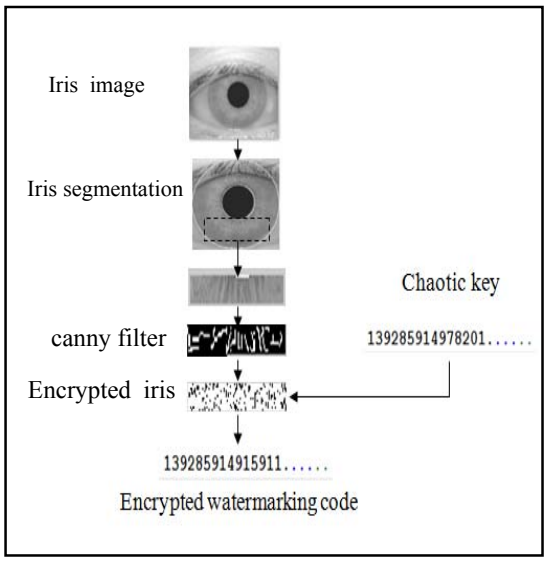


Figure 6. The proposed process for getting watermarking code

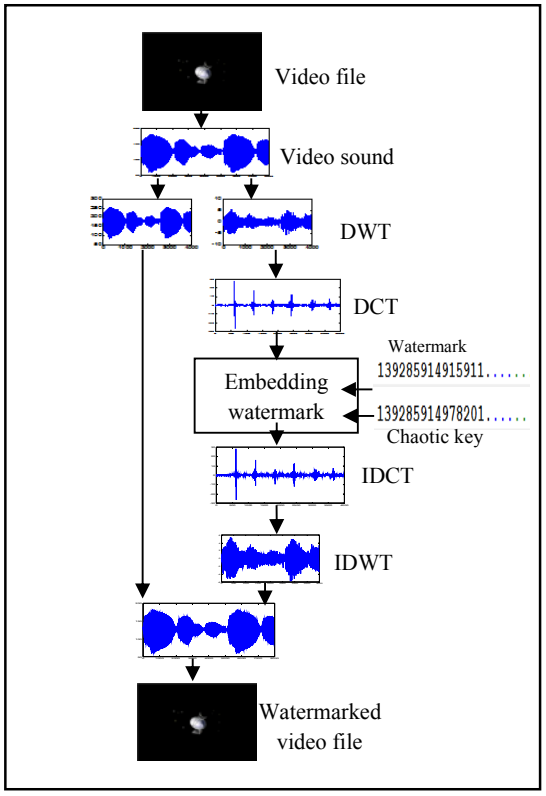


Figure 7. The proposed embedding process

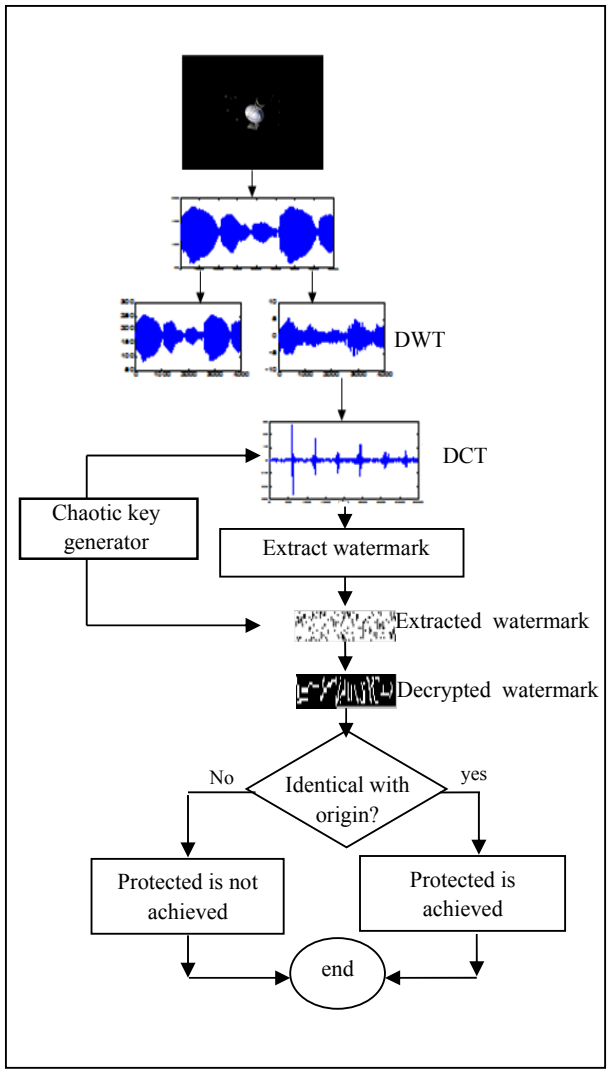


Figure 8. the proposed extracted process

A number of measures are applied on it to make sure that the proposed algorithm is strong enough to carry the watermark safely. Table I. explain the results of applying standard measures (Correlation, SNR,PSNR and MSE) to the proposed algorithm.

TABLE I. THE RESULTS OF APPLYING STANDARD MEASURES TO PROPOSED ALGORITHM

File name	Correlation	SNR	PSNR	MSE
Radar	1	219.3514	75.586	2.7631e-08
Morale	1	205.74	75.504	2.8152e-08
Test	1	212.03	75.826	2.6145e-08

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The watermarked video was attacked by simple types of watermarking attacks. This types of attacks are try to annoy the watermark by modify the whole cover without any attempt of identifying and separating the watermark [11][12]. Adding white noise (Gaussian noise) is applied to the video cover resulting from the proposed algorithm. Fig. 9 shows the effect of adding Gaussian noise to the video cover file with different signal to noise ratio values. While Table II. explains the output results of adding Gaussian noise to the video cover .

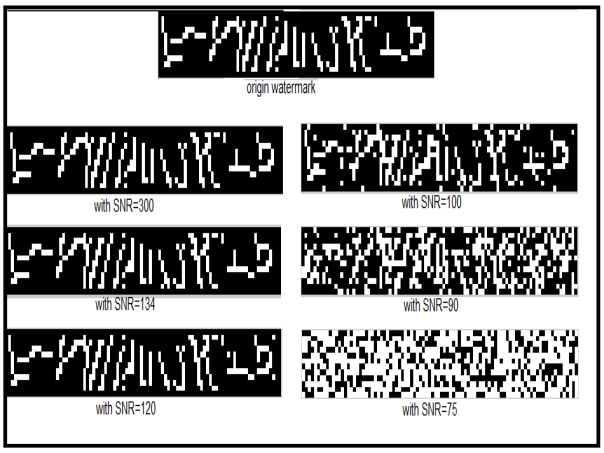


Figure 9. The effects of adding gaussian noise with variety values of signal to noise ratio

TABLE II. THE OUTPUT RESULT OF ADDING GAUSSIAN NOISE TO THE EMBEDDED WATERMARK

SNR	Correlation	MSE
200	1	0
150	1	0
134	0.8720	0.0743
120	0.7956	0.4149
100	0.1926	3.7147
90	0.0626	9.2799
75	0.0537	30.0978

VII. CONCLUSION

The paper propose an efficient method to embed a biometric watermarking in video file. It make use of two powerful mathematical transforms: DWT and DCT and applied them on the audio part of video file format instead of video's images. The proposed method use the chaotic sequence in order to find a video file locations in order to hide bio-watermark on the one hand and the sequence is used to encrypt and decrypt the bio-watermark data on the other.

After applying the proposed algorithm, some standard measures between the two watermarks(original and extracted one) are applied using correlation, SNR, PSNR and MSE. Also measures are applied on the attacked video file using correlation and MSE. The experimental results show their robustness against noise adding; very low noise watermark with expectable SNR values. The obtained results give to the proposed algorithm high performance with robustness in watermarking application in order to achieve protection to any video file.

REFERENCE

[1]. Bhaumik Arup , Choi Minkyu ,Robles Rosslin J. and Balitanes Maricel O. ," Data Hiding in Video", International Journal of Database Theory and Application, Vol. 2, No. 2, June 2009, p 9-16 .

[2]. Hood Ankita A. and Janwe N. J. ,"Robust Video Watermarking Techniques and Attacks on Watermark – A Review", International Journal of Computer Trends and Technology- volume4 Issue1 ,2013, p30-34.

[3]. Faragallah Osama S., "Efficient video watermarking based on singular value decomposition in the discrete wavelet transform domain", International Journal of Electronics and Communications (AE-) , Int. J. Electron. Commun. (AE-) 67 , 2013 , p189– 196 .

[4]. Bhatnagar Gaurav and Raman Balasubrmanian, "Wavelet packet transform-based robust video watermarking technique", Indian Academy of Sciences , Sadhana Vol. 37, Part 3, 2012, p 371–388.

[5]. Al-Gurairi Maha Abdul-Rhman Hasso," Biometric Identification Based on Improved Iris Recognition Techniques", A Ph. D. Thesis Submitted to The Council of the College of Computer and Mathematical Sciences, University of Mosul ,2006.

[6]. Waghmare L.M. and Roselin Vanaja, " Iris Texture Analysis for Security Systems" , International Journal of Computer Applications (0975 – 8887) Volume 64– No.22, 2013, p37-44.

[7]. Dhavale Sunita V. ," DWT and DCT based Robust Iris Feature Extraction and Recognition Algorithm for Biometric Personal Identification ", International Journal of Computer Applications (0975–8887) Volume 40– No.7, 2012, p 33-37.

[8] Enayatifar R. , Mahmoudi F. and Mirzaei K.," Using the chaotic map in image steganography. International Conference on Information Management and Engineering, 2009 ,p 491-495.

[9]. Saeed Melad J., " A New technique based on chaotic steganography and encryption text in DCT domain for color image", Journal of Engineering Science and Technology Vol. 8, No. 5 , 2013, p508 – 520 .

[10].Ahmed H.E., Kalash, H.M. and Farag Allah, O.S., "An efficient chaos-based feedback stream cipher (ECBFSC) for image encryption and decryption", Informatica, 31(1), 2007 ,p 121-129.

[11]. Ali Dujan Basheer Taha , "Digital Image Watermarking Techniques For Copyright Protection", A Ph. D. Thesis Submitted to The Council of the College of Computer Sciences & Mathematics , University of Mosul. , 2004.

[12].Zlomek Martin, " Video Watermarking", master thesis submitted to Department of Software and Computer Science Education, Charles University in Prague , Faculty of Mathematics and Physics, 2007.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Dynamic Cluster Formation Method To Increase The Lifetime Of A Wireless Sensor Network

K.Johny Elma
Assistant Professor, Department of IT
Jeppiaar SRR Engineering College
Chennai, India
kj_elma@rediffmail.com

Dr.S.Meenakshi
Professor and Head, Department of IT
Jeppiaar SRR Engineering College
Chennai, India
meenakshimagesh72@gmail.com

***Abstract* - A wireless sensor network consists of multiple detection stations called sensor nodes which has specialized transducers with a communication infrastructure for monitoring and recording physical and environmental conditions at diverse locations. Energy consumption of the network is crucial due to idle listening and overhearing. The sensor node's lifetime is the most critical parameter. The lifespan of a wireless sensor network is the total amount of time before the first sensor node runs out of power. An ideal cluster head is the one which has the highest residual energy. In the existing system, the cluster head loses its energy during data transmission and eventually becomes a dead node. Another node from the network is made as the cluster head. In the proposed system, we use Dynamic Cluster Formation Method to increase the lifetime of the network. In the proposed method, the clusters are formed dynamically based on its residual energy and the delay time. When the cluster head's energy drains to its threshold value, the cluster is again formed dynamically. Thus, the energy consumption is balanced by which the network lifetime is maximized.**

Keywords: Wireless Sensor Network; Sensor Node; Cluster; Residual Energy; Dead Node; Energy Consumption; Network Lifetime.

I. INTRODUCTION

A wireless sensor network (WSN) consists of sensor nodes capable of collecting information from the environment and communicating with each other via wireless transceivers. The sensor node is an autonomous small device that consists of mainly four units that are sensing, processing, communication and power supply. Sensor nodes have limited resources and it is difficult to deploy. Recharging the cluster nodes are even more difficult. Hence it is wise to use the available sensor nodes efficiently. These sensor nodes are deployed where human intervention is difficult. Hence collection of information is dependent on the sensor nodes. These sensors are used to collect the information from the environment and pass it on to base station. A base station provides a connection to the wired world where the collected data is processed, analyzed and presented to useful applications. Thus, by embedding processing and communication within the physical world, Wireless Sensor Network (WSN) can be used as a tool to bridge real and virtual environment. The collected data will be delivered to one or more sinks, generally via multi-hop communication. The sensor nodes are typically expected to operate with batteries and are often deployed to not-easily-accessible or hostile environment, sometimes in large quantities. It can be difficult or impossible to replace the batteries of the sensor nodes. Since multi-hop routing is generally needed, the nodes near a sink can be burdened with relaying a large amount of traffic from other nodes. A sensor node is a tiny device that includes four basic components. A sensing or actuating unit, a processing unit, transceiver unit and power supply unit [1, 2]. In addition to this, the sensor node may also be equipped with location detection unit such as a Global Positioning System (GPS), a mobilizer etc. Each sensing unit is responsible for gathering information from the environment as an input like temperature, pressure, light etc. and produces a related output in a form of electrical or optical signal. The analog signals produced by the sensor are converted to digital signals by the analog to digital communication (ADC) and fed into the processing unit. The transmitter and receiver are combined in to a single device called transceiver. Sensor nodes often use ISM (Industrial, Scientific and Medical) band. One of the most important components of a wireless sensor node is the power supply. The battery forms the heart of the sensor system as it decides the lifespan of the system. The battery lifespan needs to be prolonged to maximize the network lifespan. Small size of a sensor node results in corresponding constraints on memory also. Sensor nodes have very simple memory architecture.

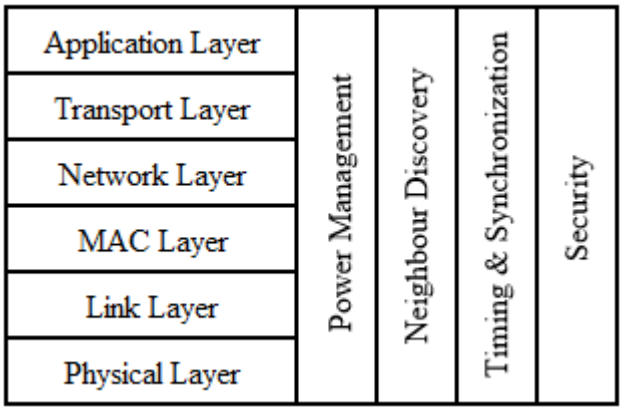


Figure 1. Protocol Stack

Sensor nodes use flash memories due to their cost and storage capacity. The mostly used operating system in sensor node are tiny OS (Operating System) Sensor nodes are resource

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

constrained in terms of energy, processor, memory, low range communication and bandwidth. Prolonging network lifetime is a critical issue. Thus, a good WSN design needs to be energy efficient. Energy consumption of one sensor node is influenced by the structure of protocol layers and the way each layer manages the sensing data.

II. RELATED WORKS

S. D. Muruganathan *et.al*, "A centralized energy-efficient routing protocol for wireless sensor network" had described the wireless sensor network and design issues [20]. They have proposed centralized routing protocol called base-station controlled dynamic clustering protocol (BCDCP), which distributes the energy dissipation evenly among all sensor nodes to improve network lifetime and average energy savings. The performance of BCDCP is then compared to clustering-based schemes such as low-energy adaptive clustering hierarchy (LEACH), LEACH-centralized (LEACH-C), and power-efficient gathering in sensor information systems (PEGASIS). Simulation results show that BCDCP reduces overall energy consumption and improves network lifetime over its comparatives.

O.B. Akan *et.al*, "Event-to-sink reliable transport in wireless sensor networks" had proposed a new reliable transport scheme for WSN, the event-to-sink reliable transport (ESRT) protocol, is presented in this paper. ESRT is a novel transport solution developed to achieve reliable event detection in WSN with minimum energy expenditure. It includes a congestion control component that serves the dual purpose of achieving reliability and conserving energy [18]. Importantly, the algorithms of ESRT mainly run on the sink, with minimal functionality required at resource constrained sensor nodes. ESRT protocol operation is determined by the current network state based on the reliability achieved and congestion condition in the network. This self-configuring nature of ESRT makes it robust to random, dynamic topology in WSN. Furthermore, ESRT can also accommodate multiple concurrent event occurrences in a wireless sensor field. Analytical performance evaluation and simulation results show that ESRT converges to the desired reliability with minimum energy expenditure, starting from any initial network state.

Wei-Peng Chen *et.al*, "Dynamic clustering for acoustic target tracking in wireless sensor networks" had devised and evaluated a fully decentralized, light-weight, dynamic clustering algorithm for target tracking. Instead of assuming the same role for all the sensors, we envision a hierarchical sensor network that is composed of 1) a static backbone of sparsely placed high-capability sensors which assume the role of a cluster head (CH) upon triggered by certain signal events and 2) moderately to densely populated low-end sensors whose function is to provide sensor information to CHs upon request. A cluster is formed and a CH becomes active, when the acoustic signal strength detected by the CH exceeds a predetermined threshold. The active CH then broadcasts an information solicitation packet, asking sensors in its vicinity to join the cluster and provide their sensing information. Through both probabilistic analysis and ns-2 simulation, we use with the use of Voronoi diagram, the CH that is usually closes to the target is (implicitly) selected as the leader and that the proposed dynamic clustering algorithm effectively eliminates contention among sensors and renders more accurate estimates of target locations as a result of better quality data collected and less collision incurred [27].

Weifa Liang *et.al*, "Online data gathering for maximizing network lifetime in sensor networks" had considered an online data gathering problem in sensor networks, which is stated as follows: assume that there is a sequence of data gathering queries, which arrive one by one. To respond to each query as it arrives, the system builds a routing tree for it. Within the tree, the volume of the data transmitted by each internal node depends on not only the volume of sensed data by the node itself, but also the volume of data received from its children. The objective is to maximize the network lifetime without any knowledge of future query arrivals and generation rates. In other words, the objective is to maximize the number of data gathering queries answered until the first node in the network fails. We then show the problem to be NP-complete and propose several heuristic algorithms for it. We finally conduct experiments by simulation to evaluate the performance of the proposed algorithms in terms of network lifetime delivered [26]. The experimental results show that, among the proposed algorithms, one algorithm that takes into account both the residual energy and the volume of data at each sensor node significantly outperforms the others.

Yong Yuan *et.al*, "Virtual mimo-based cross-layer design for wireless sensor networks" A novel multi-hop virtual multiple-input-multiple-output (MIMO) communication protocol is proposed by the cross-layer design to jointly improve the energy efficiency, reliability, and end-to-end (ETE) QoS provisioning in wireless sensor network (WSN)[28]. In the protocol, the traditional low-energy adaptive clustering hierarchy protocol is extended by incorporating the cooperative MIMO communication, multi-hop routing, and hop-by-hop recovery schemes. Based on the protocol, the overall energy consumption per packet transmission is modeled and the optimal set of transmission parameters is found. Then, the issues of ETE QoS provisioning of the protocol are considered. The ETE latency and throughput of the protocol are modeled in terms of the bit-error-rate (BER) performance of each link. Then, a nonlinear constrained programming model is developed to find the optimal BER performance of each link to meet the ETE QoS requirements with a minimum energy consumption. The particle swarm optimization (PSO) algorithm is employed to solve the problem. Simulation results show the effectiveness of the proposed protocol in energy saving and QoS provisioning.

III. SYSTEM ANALYSIS

A. Existing Model

Sensor nodes are resource constrained in term of energy, processor, memory, low range communication and bandwidth. Limited battery power is used to operate the sensor nodes and is very difficult to replace or recharge it, when the nodes die. This will affect the network performance. Energy conservation increases lifetime of the network. Wireless sensor networks

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

consist of battery-powered nodes that are endowed with a multitude of sensing modalities including multi-media (*e.g.*, video, audio) and scalar data (*e.g.*, temperature, pressure, light, magnetometer, infrared). Although there have been significant improvements in processor design and computing, advances in battery technology still lag, making energy resource the fundamental challenge in wireless sensor networks. Consequently, there have been active research efforts on performance limits of wireless sensor networks. Those operations for a sensor to consume energy are target detection, data transmission and reception, data processing, etc. Among others data transmission consumes most of the energy, and it heavily depends on the transmission distance and the amount of transmitted data.

When the data transmission occurs, the energy of the cluster head drains and eventually dies. The lifespan of a wireless sensor network is the total amount of time before the first sensor node runs out of power. LEACH is dependent on the probability model. Some cluster heads may be very close to each other. These disorganized cluster heads could minimize the energy efficiency. To overcome the defects of LEACH methodology, a cluster head selection method, High Energy First (HEF) algorithm has been introduced. This method proves that the network lifetime can be efficiently increased. For mission critical WSN applications, it is important to be aware of whether all sensors can meet their mandatory network lifetime requirements. The High Energy First (HEF) algorithm is proven to be an optimal cluster head selection algorithm that maximizes a hard N-of-N lifetime for HC-WSNs under the ICOH condition. But lifetime of the network is much lesser when compared with the proposed system.

B. *Proposed Model*

The wireless sensor network (WSN) is partitioned into several clusters based on the coverage and connectivity. First, the coverage range is checked by all the nodes in a network. This is done by broadcasting a message to all its neighbor nodes. The nodes in the sensing range send an update message to that particular node. The node which receives maximum number of messages as reply becomes a cluster head (CH). A cluster is formed based on the chosen cluster head (CH). Data transmission occurs via cluster head (CH) which means all the nodes in a cluster send their data first to the cluster head which is then passed on to the base station. From the base station, the data is being sent to the receiver. The proposed method for the project is Dynamic Cluster Formation Method (DCFM).

There are two important parameters involved in DCFM. They are the residual energy and the delay time. The node with the minimum delay time and maximum residual energy is made the cluster head. A threshold value for the energy is maintained. When the cluster head's energy drains to its threshold value, a new cluster head chosen based on its residual energy. Again, the nodes in the cluster broadcast a message to all its neighbor nodes. The nodes which are in the sensing range sends an update message to that particular node. This is done to use the energy of nodes efficiently. The cluster is again formed dynamically. Thus, the energy consumption is balanced by which the network lifetime is maximized.

IV. SYSTEM ARCHITECTURE AND PROTOCOL DESIGN

A. *System Architecture*

Dynamic Cluster Formation Method involves the formation of clusters in a network dynamically. This method is mainly used to increase the lifetime of a network which is less in the existing system. Initially all the nodes are deployed in a network. In order to divide it into clusters, the cluster head is selected.

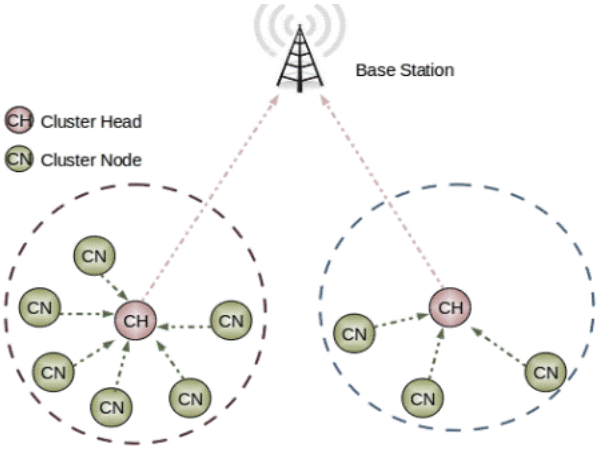


Figure 2. General Architecture of the Proposed System

The cluster head selection is dependent upon the residual energy and the delay time. The nodes which are in the sensing range of the cluster head groups to form a cluster. Initially a broadcast message is sent by all the nodes to all the other nodes. Thus, the number of broadcast messages a particular node receives is determined as node count. From the node count, the delay time is calculated. The node with minimum delay time and maximum residual energy is made the cluster head. From the cluster head the cluster is formed.

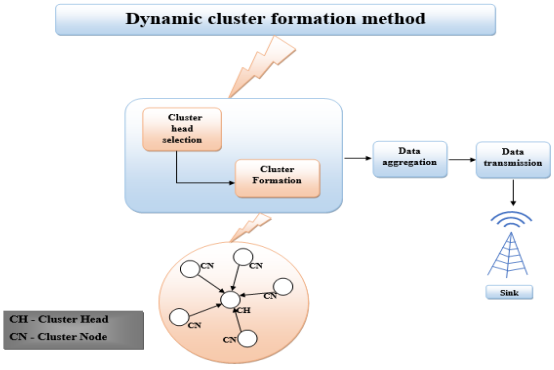


Figure 3. Proposed System Architecture

A threshold value is set for the energy of a node. During the data transmission, the energy of all the nodes drains. When

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the energy of the cluster head (CH) drains to the threshold value, another node is made the cluster head based on the residual energy and the delay time. Thus, the clusters are again formed dynamically by changing the cluster heads (CH). The cluster node (CN) senses the data and the sensed data is sent to the cluster head (CH). The cluster head (CH) receives all the data from all the nodes and does data aggregation. Once all the data are aggregated, the cluster head (CH) sends the data to the base station (BS). From the base station, the data is moved to the corresponding cluster's cluster head (CH). The cluster head (CH) transmits the data to the corresponding cluster node. Thus, the lifetime of the network is increased as there is a uniform consumption of energy in the network.

B. Protocol Design

The number of nodes is a network is denoted by the parameter 'n'. Each node has an initial energy $E_i(x)$, data transmission power $P_{tx}(x)$ and data reception power $P_{rx}(x)$. The cluster head CH is responsible for the transmission of the data from a particular cluster to other nodes. The selection of cluster head CH is dependent on the delay time $T_{delay}(x)$ and the residual energy $E_{res}(x)$. The node with the highest residual energy and lowest delay time $T_{delay\text{-}min}(x)$ becomes the cluster head CH.

Step 1: Deploy all the nodes in a network.

Step 2: For each node x, assign the initial energy $E_i(x)$, data transmission power $P_{tx}(x)$ data reception power $P_{rx}(x)$ and transmission range.

Step 3: A broadcast message is sent by the node x to all the other nodes which are in the sensing range of the node x. The message is represented as **(bcm, x).** The number of broadcast messages a particular node receives Nc(x) is determined. Here, Nc(x) is the node count of the node x.

Step 4: Calculate the delay time for the node x with node count Nc(x) as input. The delay time is given as,

$$\mathbf{T_{delay}(x) = C\ e^{1/Nc(x)}} \qquad \mathbf{(1)}$$

where C is a constant.

Step 5: The steps are repeated for all the nodes.

Step 6: The node with the lowest delay time $T_{delay\text{-}min}$ is determined from the delay time of all the nodes. An update message is sent by the node with lowest delay time to all nodes under its sensing range that it is the cluster head CH and forms a cluster. The message is represented as **(upm, x)**. If the delay time $T_{delay}(x)$ is same for more than one node, the node with the highest residual energy $E_{res}(x)$ is made the cluster head CH.

C. Energy Calculation

The energy of a node drains whenever there is a transmission or reception of data. When the energy of a cluster head CH drains to a threshold level (Thresh), another node is made the cluster head CH by following the above steps. The energy consumption of a node $E_{cmp}(x)$ is determined by the formula

$$\mathbf{Ecmp(x) = [Ptx(x) * N(tx)] + [Prx(x)*N(rx)]} \qquad \mathbf{(2)}$$

where $P_{tx}(x)$ is the data transmission power,
$P_{rx}(x)$ is the data reception power,
N(tx) is the number of transmissions,
N(rx) is the number of receptions.

The residual energy of a node $E_{res}(x)$ is determined by using the initial energy of the node $E_i(x)$ and the energy consumption of the node $E_{cmp}(x)$. It is given as,

$$\mathbf{Eres(x) = Ei(x) - Ecmp(x)} \qquad \mathbf{(3)}$$

Based upon this residual energy, the node with the maximum $E_{res}(x)$ is selected as the cluster head. The calculated residual energy is used in the selection of the cluster head. The node with the maximum residual energy and a minimum delay time is selected as the cluster head.

V. METHODLOGY

A. Dynamic Cluster Formation Technique

The sensor nodes are randomly distributed in a heterogeneous environment. The formation of cluster and energy efficient routing is done by the Dynamic Cluster Formation Method (DCFM).

- Cluster Formation: The sensor nodes are spatially distributed autonomous devices which are used for sensing, processing and communication purposes. All these nodes must be divided into clusters. Initially a network is divided into fields and the nodes are deployed in the network. The nodes which are used here are the sensor nodes which performs some processing, gathering information and communicating with each other. In order to communicate with each other, the nodes need to form a cluster. Thus, the formation of cluster is dependent upon the node's sensing range. The nodes which are in the sensing range of the cluster head becomes a member of the cluster. If a node is in the sensing range of one or more cluster, it becomes a member of the cluster which senses it first. Thus, all the nodes in a network will be a member of any one cluster of that network.
- Cluster Head Selection: Once all the nodes are deployed, it is necessary to form clusters. The formation of clusters helps in the communication of sensor nodes. The formation of cluster is basically dependent on the cluster head (CH). Hence the cluster head selection is the most important part. There are two parameters which are important in the cluster head selection. The delay time and the energy of the nodes decide which node must become a cluster head (CH). Initially, all the nodes in a network sends a broadcast message to all the other neighbor nodes. The number of message a particular node receives is determined as the node count. The delay time is calculated based upon its node count. Thus, the node with maximum energy and minimum delay time is made the cluster head. Cluster heads (CH)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

helps in the formation of clusters. Once the node with minimum delay and maximum energy is selected, the node sends an update message to all its neighbor nodes that it is the cluster head. Thus, all the nodes which receives the update message becomes a part of that particular cluster.

```
INPUT: Transmission_Power Ptx
       Reception_Power Rtx
       Initial_Energy Ei
BEGIN:
For_each_node(Current_node C(x))
{
    //Initialize Ei(x) = Ei;
    For_each_round(Current_trip C(r))
    {
        //Calculate Node_Count Nc(x);
        Calculate Transmission_Count N(tx);
        Calculate Reception_Count N(rx);
        Calculate Energy_Consumption_Of_Node Ecmp(x) with N(tx), N(rx), Ptx, Prx;
        Calculate Residual_Energy Eres(x) with Ei(x), Ecmp(x);
        //Calculate Delay_Time Tdelay(x);
    }
    Update Ei(x) with Eres(x);
}
```

The number of nodes is a network is denoted by the parameter 'n'. Each node has an initial energy Ei(x), data transmission power Ptx(x) and data reception power Prx(x). The cluster head CH is responsible for the transmission of the data from a particular cluster to other nodes. The selection of cluster head CH is dependent on the delay time Tdelay(x) and the residual energy Eres(x). The node with the highest residual energy and lowest delay time Tdelay-min(x) becomes the cluster head CH. Deploy all the nodes in a network initially, For each node x, assign the initial energy Ei(x), data transmission power Ptx(x), data reception power Prx(x) and transmission range. A broadcast message is sent by the node x to all the other nodes which are in the sensing range of the node x. The message is represented as (bcm, x). The number of broadcast messages a particular node receives Nc(x) is determined. Here, Nc(x) is the node count of the node x. Calculate the delay time for the node x with node count Nc(x) as input. The steps are repeated for all the nodes. The node with the lowest delay time Tdelay-min is determined from the delay time of all the nodes. An update message is sent by the node with lowest delay time to all nodes under its sensing range that it is the cluster head CH and forms a cluster. The message is represented as (upm, x). If the delay time Tdelay(x) is same for more than one node, the node with the highest residual energy Eres(x) is made the cluster head CH.The residual energy of a node Eres(x) is determined by using the initial energy of the node Ei(x) and the energy consumption of the node Ecmp(x).Based upon this residual energy, the node with the maximum Eres(x) is selected as the cluster head. Thus, the cluster head is dynamically selected using Dynamic Cluster Formation Method (DCFM).

- Dynamic Cluster Formation: Initially, the clusters are formed based upon the delay time and energy. Thus, the data transmission is involved where by which the energy of the nodes reduces. The energy is lost during the data transmission as well in the data reception. All the nodes in a cluster sends their data only via their cluster head. Thus, the cluster head loses more amount of energy. In the existing system, the energy drains completely and the node eventually dies. But this is the major disadvantage which is present in the existing system. The network lifetime is also very less. In order to overcome this disadvantage, we propose a method called Dynamic Cluster Formation Method (DCFM). By following this method, the network lifetime is increased. We use a threshold value for the energy of the node. This threshold value is used to balance the energy of the nodes in a network. The energy of the node decreases in different phases. They lose their energy during the data transmission and also in the data reception. Since cluster head is involved in the data aggregation, it receives the data from all the nodes. Thus, there would be a greater loss of energy in the cluster head. When the energy of the cluster head drains to the threshold value, a new node is made the cluster head based on DCFM. This is done by choosing a node which has maximum residual energy and minimum delay time. The same process is repeated on the loss of energy at the threshold level. For calculating the residual energy, the energy consumption is calculated. The energy consumption is calculated by taking into account the data transmission power Ptx(x), data reception power Prx(x), number of transmissions N(tx) and number of reception N(rx). Thus, by multiplying the number of transmissions and the transmission power along with the number of receptions and the reception power, the energy consumption is being determined. From the energy consumption, the residual energy is determined. Thus, the clusters are formed dynamically with the help of their residual energy.
- Energy Efficient Routing: In contrast to simply establishing correct and efficient routes between pair of nodes, one important goal of a routing protocol is to keep the network functioning as long as possible. The goal can be accomplished by minimizing cluster node's (CN) energy not only during active communication but also when they are inactive. Transmission power control and load distribution are two approaches to minimize the active communication energy, and sleep/power-down mode is

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

used to minimize energy during inactivity. The parameters which involves in energy consumption include,

- Time to partition a network,
- Variance in node power levels,
- Cost/packet
- Maximum node cost.

The first metric is useful to provide the min-power path through which the overall energy consumption for delivering a packet is minimized. Here, each wireless link is annotated with the link cost in terms of transmission energy over the link and the min-power path is the one that minimizes the sum of the link costs along the path. However, a routing algorithm using this metric may result in unbalanced energy spending among mobile nodes. When some nodes are unfairly burdened to support many packet-relaying functions, they consume more battery energy and stop running earlier than other nodes disrupting the overall functionality of the ad hoc network. Thus, maximizing the network lifetime (the second metric shown above) is a more fundamental goal of an energy efficient routing algorithm: Given alternative routing paths, select the one that will result in the longest network operation time. The routing protocol that is used here is Ad-hoc on-demand Distance Vector Routing (AODV).

- Adhoc On-Demand Distance Vector Routing: The reactive on demand routing protocols establish the route to a particular destination only if it is needed. Adhoc on-demand Distance Vector (AODV) is one of the commonly used reactive on demand routing protocols in mobile ad hoc network (MANET). AODV is a reactive enhancement of the DSDV protocol. The route discovery process involves ROUTE REQUEST (RREQ) and ROUTE REPLY (RREP) packets. The source node initiates the route requested through the route discovery process using RREQ packets. The generated route request is forwarded to the neighbors of the source node and this process is repeated till it reaches the destination. On receiving a RREQ packet, an intermediate node with route to destination, it generates a RREP containing the number of hops required to reach the destination. All intermediate nodes that participates in relaying this reply to the source node creates a forward route to destination. AODV minimizes the number of packets involved in route discovery by establishing routes on-demand. The sample15.tcl shows a node configuration for a wireless mobile node that runs AODV as its adhoc routing protocol. Prior to the establishment of communication between the source and receiver node, the routing protocol should be mentioned to find the route between them. Data Transmission is established between nodes using UDP agent and CBR traffic.

VI. RESULTS AND DISCUSSION

The main aim of the project is to improve the network lifetime. A cluster is collection of nodes and in this case, sensor nodes are grouped to form a cluster in a network. This is done by choosing the cluster head dynamically using the method DCFM. It is seen that the energy is uniformly utilized and the network lifetime is increased when compared with that of High Energy First (HEF) algorithm. It is graphically represented using XGraph. To analyze a particular behavior of the network, users can extract a relevant subset of text-based data and transform it to a more conceivable presentation. Thus it is proven that the network lifetime is increased.

- Deployment of Nodes: Wireless Sensor Network consists of multiple sensor nodes. All the nodes are deployed in the network in such a way that the nodes can communicate with each other. The nodes senses the information and sends it to the base station. The nodes communicate with only their neighboring nodes.

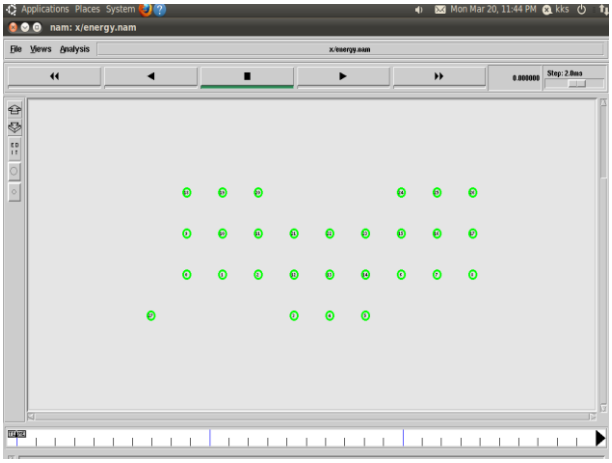
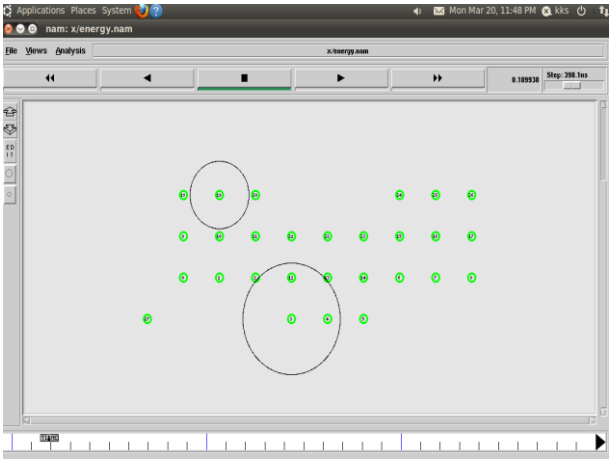


Figure 4. Deployment of Nodes

- Coverage Sensing: Identification of the nodes which are surrounding a particular node is done by sensing. A node does coverage sensing to find out its neighbor nodes for communicating. The nodes senses and sends it to the neighboring nodes and then finally it reaches the base station.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 5. Coverage Sensing

- Cluster Formation: After the coverage is sensed, all the nodes in a network will be a member of a cluster. Thus various clusters are being formed according to their sensing range. Now the transmission of data is done through the cluster head. A node is made the cluster head based upon the energy and delay time. All the nodes in a cluster transmits the data via the corresponding cluster head. The cluster head aggregates the data and passes it to the base station.

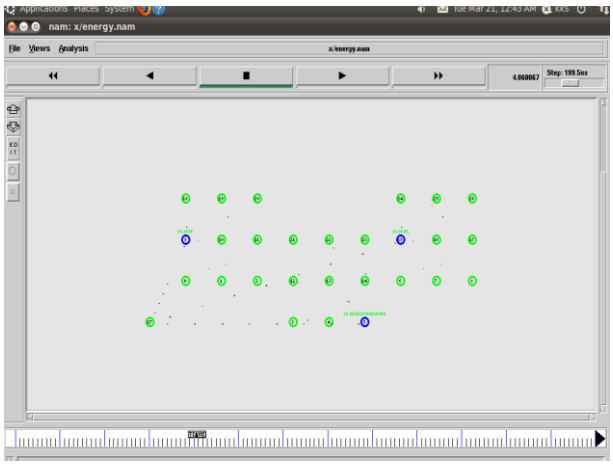


Figure 6. Formation of Clusters

- Cluster Head Selection: The nodes in a network sends a broadcast message to all the other neighbor nodes. The number of message a particular node receives is determined as the node count. The delay time is calculated based upon its node count. Thus, the node with maximum energy and minimum delay time is made the cluster head. Cluster heads (CH) helps in the formation of clusters. Once the node with minimum delay and maximum energy is selected, the node sends an update message to all its neighbor nodes that it is the cluster head. Thus, all the nodes which receives the update message becomes a part of that particular cluster.

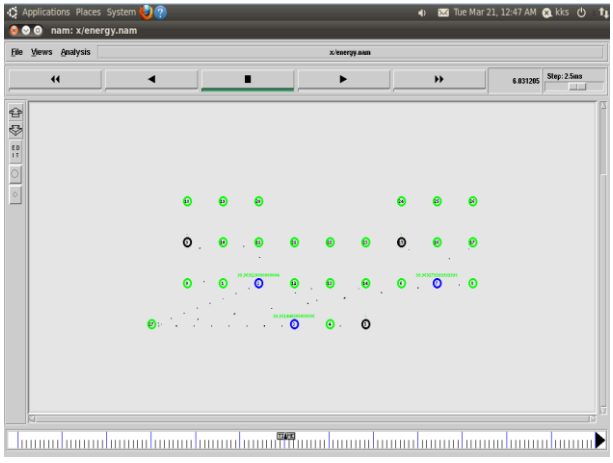


Figure 7. Cluster Head Selection

- Dynamic Clusteer Formation: All the nodes in a cluster sends their data only via their cluster head. Thus, the cluster head loses more amount of energy. We use a threshold value for the energy of the node. This threshold value is used to balance the energy of the nodes in a network. . When the energy of the cluster head drains to the threshold value, a new node is made the cluster head. Thus the node with maximum residual energy is made the cluster head.

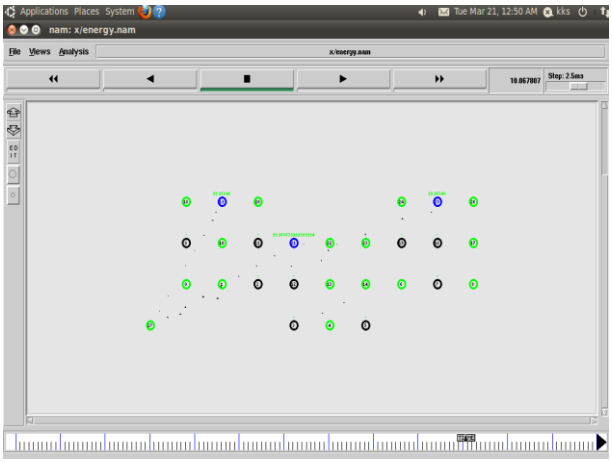


Figure 8. Dynamic Cluster Formation

Thus the energy is efficiently utilized and the lifetime of the network increases efficiently. Our experiment results show that the Dynamic Cluster Formation Method (DCFM) achieves significant performance improvement over High Energy First (HEF) algorithm, and DCFM's lifetime can be bounded.

VII. PERFORMANCE EVALUATION

The lifetime of a network is generally defined as the duration from the start to when the percentage of dead nodes comes to a threshold. It is seen that the lifetime of the network is less in High Energy First (HEF) algorithm. When compared with DCFM, DCFM provides a better network lifetime. Figure 9 has the representation of network lifetime. The red line denotes High Energy First (HEF) algorithm and the green line represents Dynamic Cluster Formation Method (DCFM). Thus, it is proved that the network lifetime is improved in Dynamic Cluster Formation Method (DCFM).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

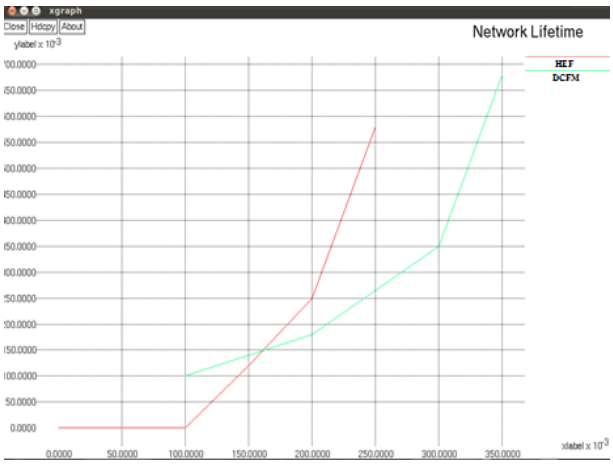


Figure 9. XGraph of Network Lifetime

- Energy of a node: Figure 10 represents the energy of a particular node. It is seen that the energy of the node is maximum at the initial state. Data transmission, data reception causes loss of energy in a node. It is seen that the node's energy drains eventually. At a particular threshold level, another node is made the cluster head. A cluster is formed only with the help of cluster head. The selection of cluster head is an important part. Thus, the selection of cluster head is dependent upon two major parameters. The residual energy and the delay time. The node with maximum residual energy and minimum delay time is made the cluster head. The delay time is calculated based on the node count. Node count is the number of broadcast message a node receives. With the node count, the delay time is calculated. From the calculated delay time, it is possible to determine the cluster head depending upon its residual energy.

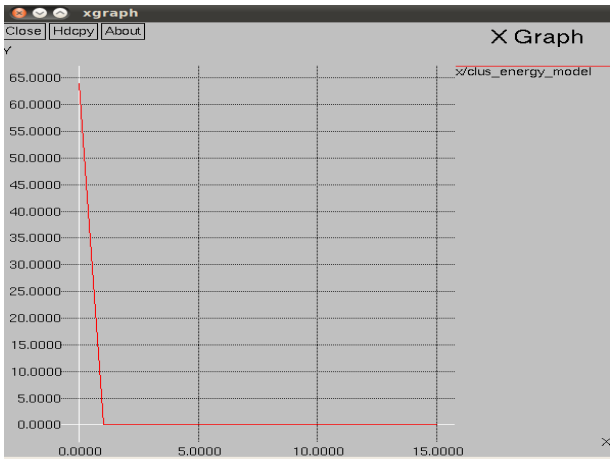


Figure 10. Energy of a Node

- Delay of a Node: Figure 11 denotes the delay of a node. The node with minimum delay and maximum residual energy is made the cluster head and it is used in Dynamic Cluster Formation Method (DCFM).

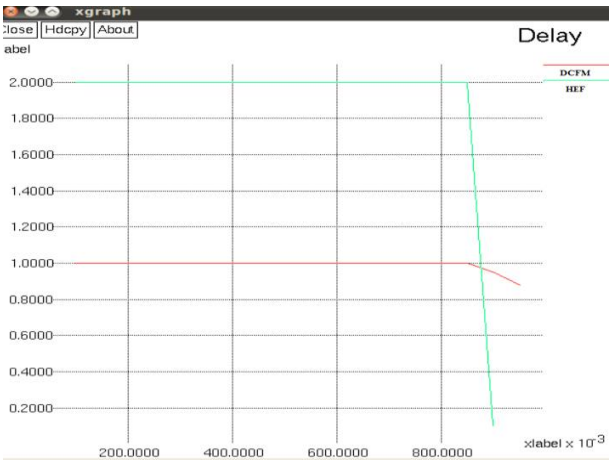


Figure 10. Delay of a Node

The residual energy and the delay time are important in evaluating the network performance. Thus the node with the minimum delay time and maximum residual energy is made the cluster head according to the threshold value. By doing this the energy is utilized efficiently thereby increasing the network performance.

VIII. CONCLUSION

On providing a trustworthy system behavior with a guaranteed hard network lifetime is a challenging task to safety-critical and highly-reliable WSN applications. For mission critical WSN applications, it is important to be aware of whether all sensors can meet their mandatory network lifetime requirements. In this project, we have addressed the issue of the predictability of collective timeliness for WSNs of interests. First, the Dynamic Cluster Formation Method (DCFM) is proven to be an optimal cluster head selection algorithm then, provide theoretical bounds on the feasibility test for the hard network lifetime. As there is an enhancement only in the network lifetime for now, there would be a greater chance of increasing the coverage and connectivity of the wireless sensor network (WSN) with a balanced energy consumption and an increased network lifetime.

REFERENCES

[1] Aurenhammer. F, "Voronoi Diagrams-A Survey of a Fundamental Geometric Data Structure," ACM Computing Surveys, vol. 23, pp. 345-405, 1991.

[2] Bandyopadhyay. S and Coyle. E.J, "An Energy Efficient Hier- archical Clustering Algorithm for Wireless Sensor Networks," Proc. IEEE INFOCOM 2003, Apr. 2003.

[3] Bertsekas. D.P, Nonlinear Programming, second ed. Belmont, Mass.: Athena Scientific, 1999. J. Bacon, D.M. Eyers, J. Singh, and P.R. Pietzuch, "Access Control in Publish/Subscribe Systems," Proc. Second ACM Int'l Conf. Distributed Event-Based Systems (DEBS),2008.

[4] Chen. H, C.-S. Wu, Y.-S. Chu, C.-C. Cheng, and L.-K. Tsai, "Energy Residue Aware (ERA) clustering algorithm for leach- based wireless sensor networks," in Second International Conference on Systems and Networks Communications (ICSNC 2007).

[5] Dongyao Jia, Huaihua Zhu, Shengxiong Zou and Po Hu., "Dynamic cluster head selection method for wireless sensor network," IEEE Sensor Journal, June 2015.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[6] Ghiasi et al. S, "Optimal Energy Aware Clustering in Sensor Networks," MDPI Sensors, vol. 2, no. 7, July 2002.

[7] Govindan. R, Hellerstein. J.M, Hong. W, "The Sensor Network as a Database," Technical Report 02-771, Computer Science Dept., Univ. of Southern California, Sept. 2002.

[8] Hansen. E, Neander. J, Nolin. M, and Björkman. M, "Energy-efficient cluster formation for large sensor networks using a minimum separation distance," in In The Fifth Annual Mediterranean Ad Hoc Networking Workshop, Lipari, Italy, June 2006.

[9] Hu Bin,Chen Wen. Analysis of Wisdom Agricultural Systems based on Wireless Sensor Network[J]. Sensors and actuators chemical,2013,(8):57.

[10] Intanagonwiwat. C, Estrin. D, Govindan. R, and Heidemann. J, "Impact of Network Density on Data Aggregation in Wireless Sensor Networking," Proc. 22nd IEEE Int'l Conf. Distributed Computing Systems, pp. 457-458, 2002.

[11] Javaid N, Waseem M, Khan Z A. ACH: Away Cluster Heads Scheme for Energy Efficient Clustering Protocols in WSNs[C]. 2013 Saudi International Electronics, Communications and Photonics Conference, Piscataway: IEEE.2013:364-7.

[12] Johnson. D, Maltz. D, and Hu. D, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)," IETF Internet draft, Feb. 2002.

[13] Kang.I and Poovendran. R, "Maximizing Static Network Lifetime of Wireless Broadcast Ad Hoc Networks," Proc. IEEE Int'l Conf. Comm. (ICC '03), 2003.

[14] Kaur. H and Baek. J, "A strategic deployment and cluster-header selection for wireless sensor networks," IEEE Transactions on Consumer Electronics, vol. 55, no. 4, November 2009.

[15] Liu H. Energy-balanced clustering algorithm for wireless sensor networks (EBCA) [J]. Electronic Measurement Technology, 2015.

[16] Mainwaring. A, Polastre. J, Szewczyk. J, Culler. D, and Anderson. J, "Wireless sensor networks for habitat monitoring," in Proc. WSNA, Atlanta, Georgia, Sept. 2002.

[17] Min. R, Bhardwaj. M,Cho et al.S. H, "Low-power wireless sensor networks," in VLSI Symp. Tech. Dig., Jan. 2001.

[18] O.B. Akan *et.al*, "Event-to-sink reliable transport in wireless sensor networks".

[19] Raghunathan. V et al., "Energy-Aware Wireless Microsensor Networks," IEEE Sig. Proc. Mag., vol. 1, no. 2, Mar. 2002.

[20] S. D. Muruganathan *et.al*, "A centralized energy-efficient routing protocol for wireless sensor network".

[21] Shah R C, Rabaey J M. Energy aware routing for low energy ad hoc sensor networks[C]. Prof.Of the 3rd IEEE Wireless Communications and Networking Conf.(WCNC),Orlando,2013:151-165.

[22] The Network Simulator – NS-2, http://www.isi.edu/nsnam/ns/

[23] Tyagi S, Kumar N. A systematic review on clustering and routing techniques based upon LEACH protocol for wireless sensor networks[J]. Journal of network and computer applications: 2013,36(2):623-645.

[24] Wang Sheng-Shih, Chen Hung-Chang, Chen Ze-Ping " An Energy and Link Efficient Clustering Technique for Reliable Routing in Wireless Sensor Networks"(IEEE), 2012.

[25] Wang.W and Jantsch. A., "An algorithm for electing cluster heads based on maximum residual energy," in Proceeding of the 2006 international conference on Communications and Mobile Computing, July 03-06, 2006.

[26] Wei-Peng Chen et.al, "Dynamic clustering for acoustic target tracking in wireless sensor networks".

[27] Weifa Liang et.al, "Online data gathering for maximizing network lifetime in sensor networks".

[28] Yong Yuan *et.al*, "Virtual mimo-based cross-layer design for wireless sensor networks".

Context based Power Aware Multi-Effector Action optimized Reinforcement Learning

Mr. S. Senthil Kumar, Research Scholar. Dr. T. N. Ravi, Assistant Professor, Periyar EVR College

Abstract*— Multi-Effector Action Optimized Reinforcement Learning provides a configurable intruder detection system with dynamic security procedure switching schemes using one of the best Machine Learning (ML) procedures Reinforcement Learning (RL). An automated 'security threshold determining procedure' based on the active heterogeneous network circumstances is provided here to operate with Reinforcement Learning in the name of "Context based Power Aware Multi-Effector Action optimized Reinforcement Learning" (CPAMEA-RL). This procedure finalizes the security threshold values based on the context of the data. This value is important to choose an optimum security scheme which works on pre-calculated computational-power guidelines, so that the network security administration is provided with amended power utilization.

***Index Terms*— Reinforcement Learning (RL), Machine Learning (ML), Multi-Effector Action optimized Reinforcement Learning (MEA-RL), Context based Power Awareness, Security threshold determining, security based on computational-power guidelines.**

I. Introduction

Modern communication mostly carried out by a number of clusters of mixed type electronic network nodes. This heterogeneous network communication has a wide range of data and bandwidth utilization. Communication protocols and security policies among the cluster nodes are mostly diversified based on the nodes categories. Most of the nodes are battery operated or rechargeable at least and they equipped with beneficial mobility. This precariousness nature of nodes makes the clusters dynamic and causes the entire network into a less predictable entity. Providing the best security for this network without affecting its Quality of Service (QoS) is a challenging job. The QoS of any network is depending on the standard network parameters of Throughput, Communication Delays, Level of Security, and Power consumption.

Increasing the positive factors of QoS like throughput, security while decreasing the negative factors like jitter, latency, End-to-End delay and Power consumption is the vital aspiration while designing a raw network architecture. Intervening manually entire communication of this modern network pattern consumes more computational resources whereas the results are not up to the mark. This situation makes the manual security monitoring as a desolate task. Providing automatic security to this network improves the QoS because of the modern Machine Learning and Artificial Intelligence procedures. A substantially good Network Simulator like OPNET can be used to create a replica of the real world modern heterogeneous network along with existing and proposed security models the benchmark parameters like throughput, jitter, latency, end-to-end delay, security and power consumption can be measured. A number of simulations are performed with random network node placements and with random communications the benchmark results are measured and tabulated. These tabulated values are used to calculate the significance level of QoS improvement by using statistical calculations.

II. Related Works

A number of automated security policy selection schemes are contrived in the past decade. The major classification of these schemes based on Artificial Intelligence [1], Neural Network [2], Machine Learning [3] and Data mining technologies [4]. Some of the hybrid security policy schemes are running based on combining multiple technologies in simultaneous or sequential mode. Based on the standard network QoS parameters, upmost qualified technologies are selected into comparison.

The selected procedures are

1. Reinforcement Learning (RL)
2. Reinforcement Learning based on MDP (RL-MDP)
3. Multi-Effector Action Reinforcement Learning (MEA-RL)

1. Reinforcement Learning (RL):

Reinforcement Learning overcomes the disadvantages of its former procedures and performed well with dynamic independent data. RL combines both active and passive approach learning simultaneously. RL adopts the natural

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This **work** was supported in part by the U.S. Department of Commerce under Grant BS123456".

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

learning method of midbrain dopamine that learns by performing reward oriented prediction. Each knowledgebase entry of RL resembles the actual firing of a dopamine neuron. RL periodically updates its knowledgebase based on 'State-Action and Reward State-Action'. These knowledgebase updates are used to effectively train the system even with independent data. Thus Reinforcement Learning is recommended to introduce Artificial Intelligence (AI) based network security.

RL considers all 41 essential network security data factors from KDD-Cup dataset [5]. They are duration of the connection, type of the communication, service type, communication flag, Number of source data Bytes, Number of destination data Bytes Geographical location, Fragment Errors, Priority, Communication mode, Number of failed logins, Login flag, compromised network connections, Root access, Number of root access attempts, Number of file creations, Number of shell access, Number of files accessed, Number of outbound commands, Host Login flag, Guest Login flag, Login counts, Service Count, Service error count, connection error rate, service discard rate, guest-host service ratio, guest-host differential service rate, Service differential ratio, Number of destination hosts, Destination host service count, Destination host same service rate, Destination host different service rate, Destination host port match rate, Destination host server different host rate, Destination host server error rate, Destination host server service error rate, Destination host error rate, Destination host server recent error rate.

All these parameters are involved in calculating decision making factors for RL. Expected sum of immediate and long time rewards under the more suitable policy referred as Utility. It is calculated as

$$util(s_t, a) = E\left\{\text{Reward}(s_t, a) + \max policies \sum_{j=1}^{N-1} R_{t+j}\right\}$$

Where s_t refers the state at particular timestamp t, $\text{Reward}(s_t, a)$ refers immediate reward of executing action a in state
s_t, N refers number of steps taken by the agent in its lifetime. E{.} refers expectation over all possible combination of decisions.

Sometimes RL abides by taking reward oriented heuristic decisions makes the security system vulnerable to strategic long term attacks. In this criterion RL needs larger time consuming updates in its knowledgebase which makes the security system less responsible to the real-time data.

2. Reinforcement Learning with Markov's decision Process (RLMDP):

Knowledge base updates in RL are consuming more time against 'strategic attacks' and this problem is solved in RLMDP. Markov's Decision Process reduces many inutile heuristics movements performed in RL. Whenever there is an ambiguous decision or a decision with less support count occurs, RL took the default action expecting a reward whereas RLMDP applies MDP and filters the action if there is a less probability to get the reward. This nature of RLMPD makes it more stable against different attacks.

Markov Decision Processes (MDPs)[6][7] are operates on high dimensional state and action spaces represented as s and a respectively. To get the state s_t, action a_t and reward r_t at time t the state transition combinational probabilities and expected reward function is declared as $P(s_{t+1}|s_t, a_t)$ and $R(s_t, a_t)$. Stochastic and stationary polices declared by conditional distributions over actions $\pi^\theta(a; s)$ parameterized by θ. It is assigned that given policy π^θ the MDP is ergodic with stationary distribution d^θ. In RLMDP energy-based policies are considered which can be expressed as conditional joint distributions over actions a and a set of latent variables h

$$\pi^\theta(a, n; s) = \frac{1}{Z(s)} e\phi(s, a, h)\top_\theta \rightarrow (1)$$

where *φ(s,a,h)* are a pre-defined set of features and $\Sigma_{a,h} exp(e\phi(s, a, h)\top_\theta)$ is the normalizing partition function. The policy itself is then obtained by marginalizing out h. Latent type variables used to make policies based on energy and these policies classify composite non-linear and non-product relationship between actions and states inherent classification (1) is log linear with the features *φ(s,a,h)*. In a conditional restricted Boltzmann machine (RBM), the states *s*, actions a and latent variables h are all high dimensional binary vectors, and (1) is parameterized as

$$\pi^\theta(a, n; s) = \frac{1}{Z(s)} e^{s\top w_s h + a^\top w_a h + b_s\top s + b_h\top h + b_a\top a} \rightarrow (2)$$

where the parameters are matrices *Ws, Wa* and vectors *bs, ba, bh* of appropriate dimensionalities. Marginalizing out *h*, used to get a non-linearly parameterized policy

$$F^\theta(s, a) = -b_s\top s - b_a\top a - \Sigma_i \log\left(1 + e^{s^\top w_{si} + a^\top w_{ai} + b_{hi}}\right),$$
$$\pi^\theta(a; s) = \frac{1}{z(s)} e^{-F(s,a;\theta)} \rightarrow (3)$$

where *i* indices the latent variables, and w_{si}, w_{ai}, b_{hi} are parameters associated with latent variable h_i. The quantity $F(s, a; \theta)$ is the conserved energy.

The policy selection is constantly updated by SARSA, the state action pairs can be the nearest neighbor nodes. Here physical position of cluster information is used instead of Virtual Power Cluster (VPC) to reduce computational power. The error rate of SARSA can be computed as

$$\varepsilon(s^t, a^t) = [r^t + \gamma Q(s^{t+1}, t^{t+1})] - Q(s^t, a^t) \rightarrow (4)$$

In case the state-action function is determined by θ, then the update equation for new parameter is

$$\Delta\theta \propto \varepsilon(s^t, a^t)\nabla_\theta Q(s^t, a^t) \rightarrow (5)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The update process determines the security policy $M_x(R)$for the corresponding cluster P_x. The RL system was pre-trained to a beginning level with the optimal action

$$P(a|s) \approx \frac{e^{Q(s,a)/\tau}}{z} \rightarrow \quad (6)$$

where Z is a normalization factor, τ is a positive number represents the iteration. The RL convergence can be identified with the value of τ, if it is getting higher values, then it refers the RL training is under progress is with uniform improvement.

RLMDP operates with more power awareness than the other existing methods discussed here. The number of battery operated nodes is increasing in modern network. Therefore providing more security with less power consumption is important in modern network security systems. The concept of Virtual Power Clusters (VPC)[8][9] is used in RLMDP to facilitate a balanced action between power and security. The lack of parallelism and linear State Action - Reward State Action are main disadvantages of RLMDP and this affects the performance of RLMDP while dealing real-time data.

3. Multi-Effector Action Reinforcement Learning (MEA-RL):

Multi-Effector Action Reinforcement Learning (MEA-RL) is designed to use parallel sandboxing technique. Two sandboxed environments are set up in each cluster head to monitor all incoming connection requests. State action – Reward State action are performed in parallel based on the 41 network parameters of the incoming connection request and the reward attaining decision is updated in the knowledgebase. The decision which is miscarried the reward is inflicted in the knowledge base.

In MEA-RL, if $P(s)$ is a decision making policy with any of the mapping from states to actions, then the policy action quotient Q^P can be calculated as

$$Q^P(s_t, a_t) = E[r1+\gamma r_{i+1} + \gamma^2 r_{i+2} + \gamma^3 r_{i+3} \ \ldots | s_i, a_i, p]$$

Futures states can be calculated by performing recursive form as

$$Q^P(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}|s_t, a_t)\, Q^P(s_{t+1}, P(s_{t+1}))$$

Optimal value function along with associated policy can be calculated as

$$Q^*(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}|s_t, a_t) \max_{a_{t+1} \in A} Q^*(s_{t+1}, a_{t+1})$$

Consider the mean-estimate rule is $\mu_k(s_k)$, then error driven mean-estimation is calculated using

$$\mu_{k+1}(s_k) = \mu_k(s_k) + k_k \cdot \delta_k$$

where k_k is knowledge base update rate (learning rate) and it is calculated using

$$k_k = \frac{\sigma_k^2(s_k)}{\sigma_k^2(s_k) + \sigma_r^2(s_k)}$$

Prediction error is calculated using $\delta_k = \gamma_k - \mu_k$

III. Proposed Method & Implementation

Context based Power Aware Multi-Effector Action optimized Reinforcement Learning consists of 3 main concepts. CPAMEA-RL follows IPv6 protocols header format added with additional contents to incorporate the following concepts.

1. Data Sensitivity
2. Sensitivity Bits
3. Security Protocol Allocator

1. Data Sensitivity:

Modern Network data is an aggregation of multiple sensitive data. Each data has its own importance and sensitivity. Data Sensitivity module of CPAMEA-RL is designed to operate based on the data sensitivity regulation [10] recommended by Massachusetts Institute of Technology (MIT). As per the guidelines, data are classified in four sensitivity threshold. The highest security index data are Credit Card and Bank Account details, Social Security Numbers, Personal Medical Data, Military related documents and confidential data of Research Organizations. This kind of data should be kept confidential and should be handled with proper security authentications.

The second sensitivity category is high confidential index data like financial information, information disclosed by non-disclosure agreements, management information and contract details. These data are containing a highest security request tag in general, where there are two possibilities of security services subsists. If the host has enough power to process these data with high security authentication, then these data are treated like high security index data and security is gained by the highest authentication procedures. Another case, if the host is not having sufficient power to process these data with highest security protocols, second grade security protocols are followed to conserve power. In this case both power saving and highest security are not guaranteed but either one is assured.

The third sensitivity category is low confidential index data like social media forwarded data, public chat information, details of a shared or public library and discussion forums. These data are processed with low power consuming security procedures. Some amount of power saving is assured while handling these information in CPAMEA-RL. The fourth sensitivity category is a secure-free type data like public entertainment broadcasting data, streaming entertainment data and open libraries. These data are meant to

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

be prepared to reach almost every node without any authentication requirement from the sender side. The host has the rights to block these data individually by marking as spam or permitting these data in which power is used only for communication and not for any security crypto procedure. CPAMEA-RL is designed to handle all these four kinds of data sensitivity in a desirable mode.

2. Sensitivity Bits:

Sensitivity bits are used to mark the sensitivity of the data. Two bits are used here since there are four sensitivity classifications in CPAMEA-RL.

00 – is used for security free communication data
01 – Low security index data
10 – High security index data
11 – Highest security index data

These two bits are added as the extension header with the standard 40 Bytes IPv6 data header [11]. Structure of IPv6 standard header is given in Table 1.

S.No	Bits	Description
1	4	Version
2	8	Traffic Class
3	20	Flow Label
4	16	Payload Length
5	8	Next Header
6	8	Hop Limit
7	128	Source Address
8	128	Destination Address

[Table 1]

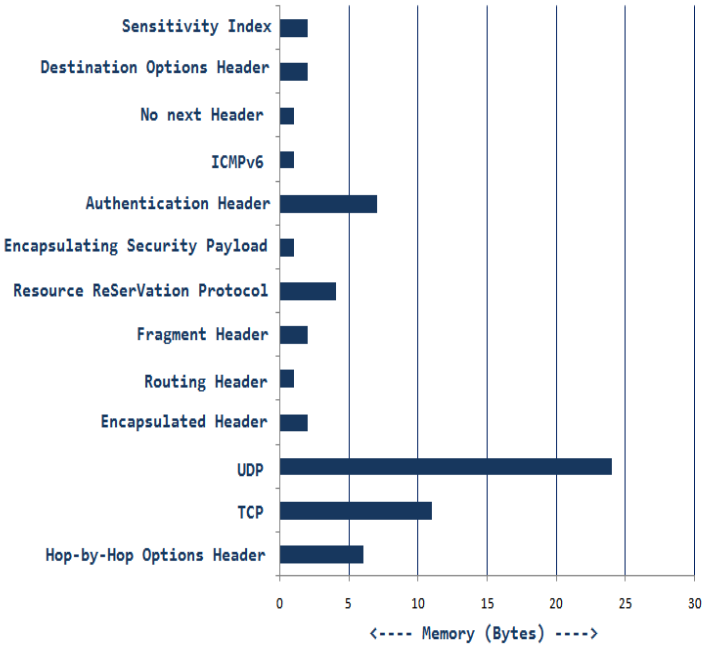
The standard extension header of IPv6 is given in table 2

Extension Header	Next Header Value	Description
Hop-by-Hop Options header	0	read by all devices in transit network
Routing header	43	Contains methods to support making routing decision
Fragment header	44	Datagram fragmentation parameters
Destination Options header	50	read by destination devices
Authentication header	51	Authenticity Information
Security Payload header encapsulation	60	Destination Options Header

[Table 2]

Sensitivity bits are added after destination options header in bit positions 61 and 62. IPv6 has similar information that assigns the security payload header encapsulation in destination options header. The difference is destination options header of the IPv6 protocol is assigned by the sender and will be processed only by the receiver. The intermediate nodes and cluster heads are not processing the security header where as the sensitivity bits are designed to process by the cluster heads. Cluster heads are authorized to allocate security resources based on the sensitivity bits' values. Since sensitivity bits are added as the last header in IPv6 protocol's header sequence, the value 59 is assigned as the next header field that refers nothing follows the sensitivity bits.

CPAMEA-RL packet header is illustrated in picture 1.



[Figure 1: CPAMEA-RL Packet Header]

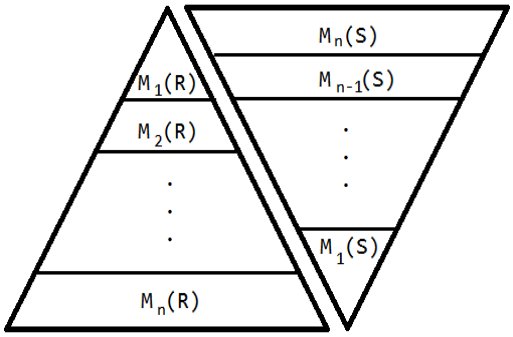
3. Security Protocol Allocator:

The actual use of Context based Power Aware Multi-Effector Action optimized Reinforcement Learning is utilized in this module. Real data sensitivities are in many layers whereas they categorized into four major types (with two reserved bits). So each major security category is consists of multiple security level layer. Allocating a suitable security protocol for each network communication that occurs in same sensitivity category with different security layer.

The Reinforcement Learning System is equipped with Memory mapping of Security Protocols. The memory map is created using two parameters they are Resource Consumption and Security Strength shown in Figure 2. In figure 1, the first triangle refers the resource consumption

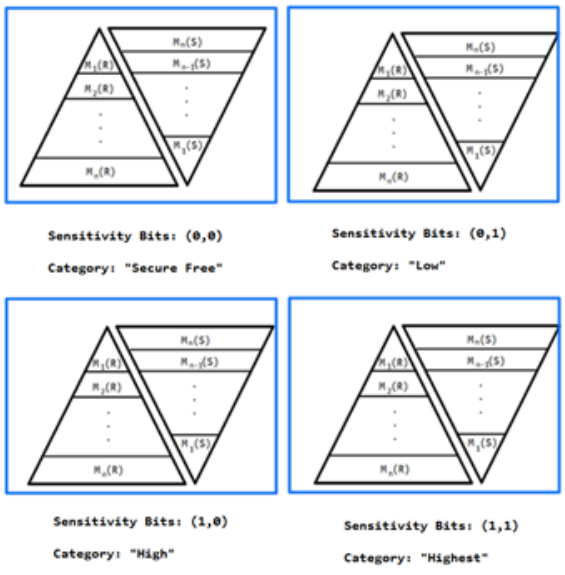
https://sites.google.com/site/ijcsis/
ISSN 1947-5500

of the security protocols from $M_1(R), M_2(R) \ldots M_n(R)$ and they are arranged in ascending order based on the resource consumption that is $M_1(R)$ consumes very less resource than any other method and $M_n(R)$ consumes the highest resource than the other methods taken in to consideration. The second triangle refers the security strengths $M_1(S)$ to $M_n(S)$ of the security protocols from $M_1(R)$ to $M_n(R)$ respectively in ascending order. That is $M_n(R)$ provides the highest security and $M_1(R)$ provides the least security while comparing with the other participating methods.



[Figure 2: Memory mapping of security protocols]

In CPAMEA-RL, the four major sensitivity categories are allocated with corresponding Memory maps with different security protocols. Each major sensitivity category uses RL to find out a desirable security protocol for the sub-security layer to involve in the communication shown in Figure 3.



[Figure 3: CPAMEA-RL Security Protocol Allocator]

The process of selecting the sensitivity category is deterministic because of the debut of sensitivity bits and selection of security protocol in a sensitivity category is non-deterministic, thus RL is applied here to solve the problem. The security protocol aggregation contains seven cryptographic procedures Rivest-Shamir-Adleman Algorithm (RSA), Data Encryption Standard (DES), Triple Data Encryption Standard (3DES), Advanced Encryption Standard (AES), Elliptic Curve Cryptography (ECC), Blowfish and International Data Encryption Algorithm (IDEA). These cryptographic procedures are configured to use different size keys based on the requirement.

AES, DES and 3DES are used predominately in Low security Sensitive Category. Low Power utilization and less computational work are involved in this sensitivity category. RSA and IDEA are used mostly in high security sensitivity category. Moderate power utilization with adequate security is achieved by using these procedures. Blowfish and ECC are used with comparatively large keys in the highest sensitivity security segment. Power is compromised here but security is the prime concern of this high secure zone.

KDD-Cup dataset is used to train CPAMEA-RL in a similar way which is used to train RL. The difference is, in CPAMEA-RL, Multi-Effector Action Optimization reduces the training time. The dataset contains 3,925,650 attacks and 972,781 normal records (4,898,431 transactions in total) are adequate to make the CPAMEA-RL.

In CPAMEA-RL, if $P(s)$ is a decision making policy with any of the mapping from states to actions and δ is the sensitivity index, then the policy action quotient Q^P can be calculated as

$$Q^P(s_t, a_t) = E[r1+\gamma r_{i+1} + \gamma^2 r_{i+2} + \gamma^3 r_{i+3} \ldots | s_i, a_i, d, p]$$

Futures states can be calculated by performing recursive form as

$$Q^P(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}, \delta_{t+1} | s_t, a_t)\, Q^P(s_{t+1}, \delta_{t+1}, P(s_{t+1}))$$

Optimal value function along with associated policy can be calculated as

$$Q^*(s_t, a_t) = r(s_t) + r \sum_{s_t+1 \in \varphi} P(s_{t+1}, \delta_{t+1} | s_t, a_t) \max_{a_{t+1} \in A} Q^*(s_{t+1}, \delta_{t+1}, a_{t+1})$$

Consider the mean-estimate rule is $\mu_k(s_k)$, then error driven mean-estimation is calculated using

$$\mu_{k+1}(s_k) = \mu_k(s_k) + k_k \cdot \delta_k$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

where k_k is knowledge base update rate (learning rate) and it is calculated using

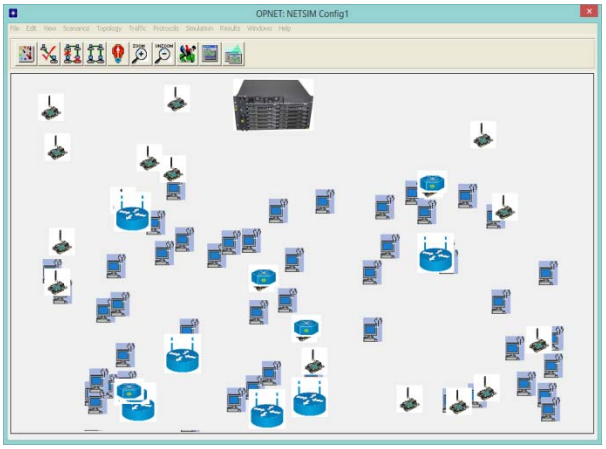
$$k_k = \frac{\sigma_k^2(s_k)}{\sigma_k^2(s_k) + \sigma_r^2(s_k)}$$

Prediction error is calculated using $\delta_k = \gamma_k - \mu_k$

IV. Performance Analysis

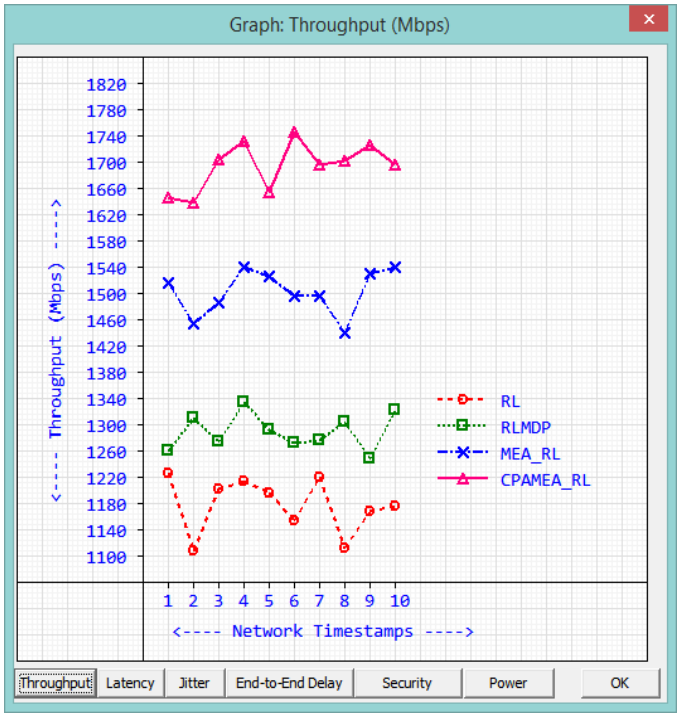
Performance of MEA-RL along with existing methods are measured by calculating the standard network QoS parameters throughput, latency, jitter, end-to-end delay, security level and average power consumption. Data are evenly distributed into four security labels ($\{\delta_0, \delta_1, \delta_2, \delta_3\}$). The data are classified into four types as Control Data (D_c), Text data (D_T), Voice data (D_v) and Multimedia data (D_m). All these types are evenly distributed in typical heterogeneous network traffic. Ten equal time stamps are selected from the simulation process. Active Reinforcement Learning, Reinforcement Learning, and Reinforcement Learning with MDP are taken as the participants in the simulation to compare with the proposed Multi-Effectors Action Reinforcement Learning. An User Interface and script wrapper to OPNET Modeler[12][13] is designed with Visual C++ programming language of Visual Studio 2013 Integrated Development Environment(IDE). NETSIMCAP – a Network Simulation and Capture Software Development Kit is used to interface Visual C++ with OPNET network simulator. Centralized Server with six Wi-Fi routers and 50 heterogeneous nodes are placed using random placement scheme[14] of OPNET to create a hybrid heterogeneous network[15][16] environment. OPNET creates required virtual network environment then process the exact network scenario while measuring the parameters instead of performing calculations. By this fact the results from OPNET are more realistic than any other calculation based results.

Example CPAMEA-RL network scenario for security enhancement: Security Level δ_x is classified as $\forall\delta_i$, i from 0 to n, $\delta_0 \leq \delta_x \leq \delta_{n-1}$ where δ_0 represents least security and δ_{n-1} represents most security. Available power resource σ_x is between least power index σ_0 to most power index σ_n are assorted in VPCs. When data with security index δ_{n-2} arrives to a VPC with power index σ_n, then δ_{n-2} will be elevated to the next security level of $\delta_{n-1}$where security is ensured. Example CPAMEA-RL network scenario for Energy efficiency [17][18]: When data with security index $\delta_1$arrives to a VPC with power index σ_0, then security index $\delta_1$will be imposed to a lesser security level δ_0. In some network transactions, this security step down activity may assailable but compromised security will not be a problem because the context is pre-assigned to a low sensitivity type by the sensitivity bits.



[Fig.5 OPNET Network structure]

Figure 5 shows the placement of heterogeneous nodes and wireless network distribution hotspots in OPNET

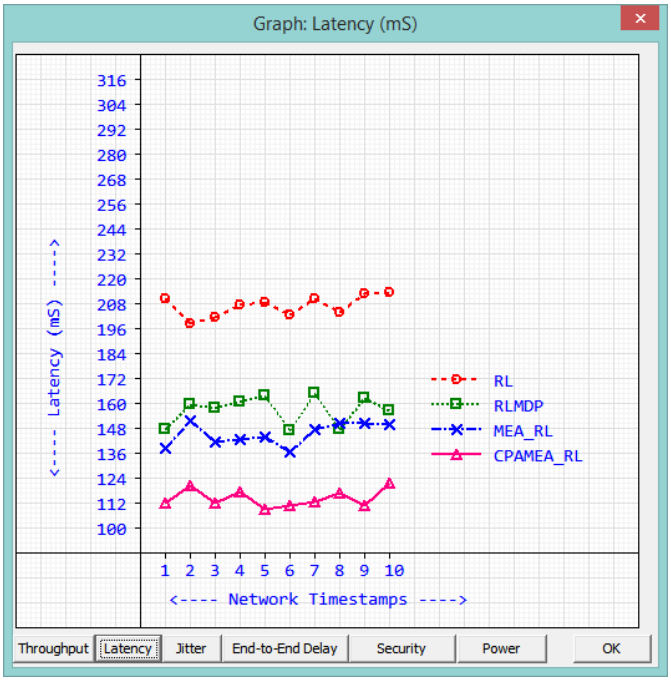


[Fig.6 Throughput Mbps]

Throughput of RL, RLMDP, MEA-RL and CPAMEA-RL are shown in Figure 6. RL achieved throughput from 1108 Mbps to 1226 Mbps based on the random locations of the nodes.

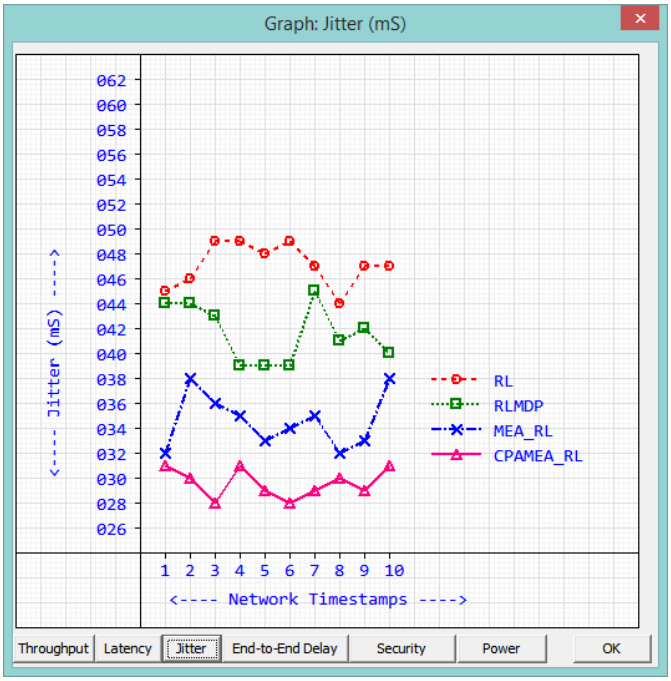
RLMDP achieved a little better than RL, i.e. from 1249 Mbps to 1334 Mbps. MEA-RL got the throughput range of 1440 Mbps to 1541 Mbps which is higher than RL and RLMDP. Whereas proposed CPAMEA-RL achieved the highest throughput of 1639 to 1746 Mbps range which is higher than all other methods taken into comparison – shown in Figure 6.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



[Fig.7 Latency (mS)]

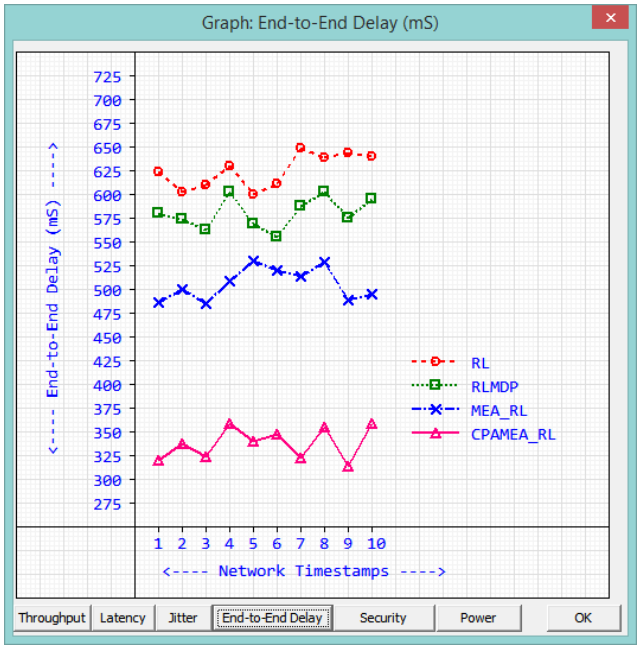
Latency is a delay in nodes response for a network transaction. This latency is inversely proportional to QoS of a network. To maintain a better QoS latency has to be kept to the bare minimum negligible level. CPAMEA-RL reduces the latency to the minimum value of 109 mS whereas other existing methods took 137 to 214 mS. Latency comparison chart of existing and proposed methods is given in figure 7.



[Fig 8. Jitter (mS)]

The time difference in packet inter-arrival time to their destination is called as jitter. Jitter is a natural delay in packet based network communication. In general, TCP and IP protocols are dealing with the jitter impact on communication. To achieve higher QoS, jitter should be kept to the minimum negligible level.

The lowest value of 28mS is achieved by the CPAMEA-RL implies the higher performance than the other methods involved. Comparison graph is shown in Figure 8.

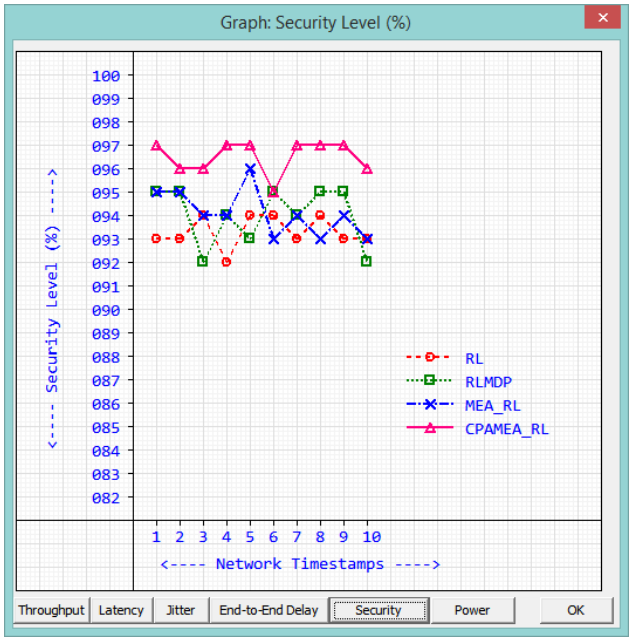


[Fig 9. End-to-End Delay]

Average travelling time taken by a data packet from source to destination is called as End-to-End Delay. It includes delays caused by route discovery process and the data packet transmission queue. Dropped packets are not considered while calculating end-to-end delay and all successfully delivered packets are included in the end-to-end delay calculation.

The measured End-to-End delay of CPAMEA-RL method is shown in Figure 9. CPAMEA-RL gets the minimum end-to-end delay of the range from 314mS to 359mS.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



[Fig.10 Security Level (%)]

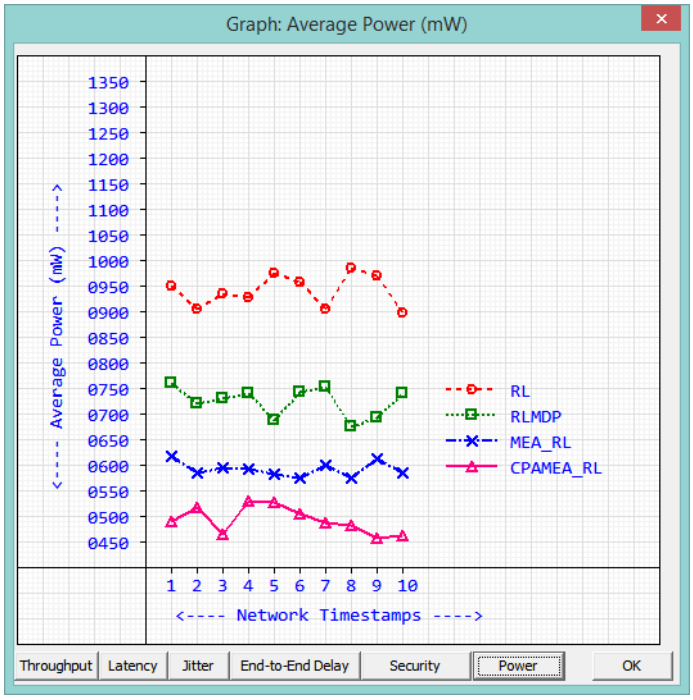
Security is the vital criteria involved in modern networks with shared and distributed infrastructures. The higher security level refers the higher quality of the network architecture. The highest security value of 97% is achieved by CPAMEA-RL is shown in Figure 10. Even though RL and RLMDP are getting closer security levels with the security level of proposed CPAMEA-RL, higher category average is achieved by CPAMEA-RL.

The security strength is measured by OPNET simulator's internal mechanism which consists of all standard attacks like Brute force attack, Dictionary attack, Wormhole attack, Sinkhole attack and etc. The measured security for MEA-RL and CPAMEA-RL are given in Table 3. CPAMEA-RL achieved 96.5% whereas MEA-RL achieved 94.1%. The improvement in security of 2.4% is a significant improvement when the security scores are above 90%.

Time stamp	MEA-RL(%)	CPAMEA-RL(%)
1	95	97
2	95	96
3	94	96
4	94	97
5	96	97
6	93	95
7	94	97
8	93	97
9	94	97
10	93	96
Average	94.1	96.5

[Table 3]

The prime target of proposed method is to provide uncompromising QoS with highest security and lowest power consumption. Based on the OPNET simulation measurements, CPAMEA-RL is used lesser power range from 459mW to 530mW. Average power consumption of existing methods with proposed methods are compared in Figure 11.



[Fig.11 Average Power (mS)]

Average power consumption is measured for MEA-RL and CPAMEA-RL is given in table 4. CPAMEA-RL used 99.5mW lesser than the MEA-RL on average. Measured power for MEA-RL and CPAMEA-RL are given in table 4.

S.No	MEA-RL(mW)	CPAMEA-RL (mW)
1	619	492
2	587	518
3	596	467
4	594	530
5	583	528
6	575	507
7	602	488
8	576	484
9	614	459
10	585	463
Average	593.1	493.6

[Table 4]

V. Conclusion

In this paper, the Reinforcement Learning method is endued with innovative Context based Power Aware Multi-Effctor Actions. Based on the observed results in a typical heterogeneous network simulation environment, CPAMEA-RL secures highest QoS indicants. The crown part of a security system is to provide highest security with lowest power consumption which is achieved in CPAMEA-RL. Since

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

CPAMEA-RL is equipped with the cutting-edge technologies, it is ready to be used in the process of constructing robust heterogeneous network environments.

REFERENCES

[1] Daniel Grzonka, Agnieszka Jakobik, Joanna Kolodziej, Sabri Pllana. Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security- Future Generation Computer Systems Elsevier June 2017

[2] Min-Joo Kang, Je-Won Kang. Intrusion Detection System Using Deep Neural Network for In-Vehicle Network Security. PLOS One Tenth Aniversary June 2016

[3] Anna L.Buczak, Erhan Guven. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials October 2015

[4] IH Witten, E Frank, MA Hall, CJ Pal. Data Mining: Practical Machine learning Tools and Techniques. Google Books 2016

[5] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A Detailed Analysis of the KDD CUP 99 Data Set – IEEE 2009

[6] Deniz Gunduz, Kostas Stamatiou, Nicolo Michelusi, Michele Zorzi. Designing intelligent energy harvesting communication systems. IEEE Communications Magazine - IEEE January 2014

[7] Mugen Peng, Shi Yan, Kecheng Zhang, Chonggang Wang. Fog-computing-based radio access networks: issues and challenges. IEEE Network Volume: 30, Issue: 4 - IEEE August 2016

[8] Ado Adamou Abba Ari, Blaise Omer Yenke, Nabila Labraoui, Damakoa, Abdelhak Gueroui. A power efficient cluster-based routing algorithm for wireless sensor networks: Honeybees swarm intelligence based approach. Journal of Network and Computer Applications - Elsevier April 2016

[9] Zhongyuan Zhao, Mugen Peng, Zhiguo Ding, Wenbo Wang, H.Vincent Poor. Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks. IEEE Journal on Selected Areas in Communications Volume: 34, Issue: 5 - IEEE March 2016

[10] Juan C. Asenjo. Data Masking, Encryption, and their Effect on Classification Performance: Trade-offs Between Data Security and Utility. Nova Southeastern University http://nsuworks.nova.edu/gscis_etd/1010. June 2017

[11] P. Thubert, Ed, C. Bormann, L. Toutain, R. Cragie. IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Routing Header. RFC Editor RFC 8138 https://www.rfc-editor.org. April 2017

[12] M. Fazeli, H. Vaziri. Assessment of Throughput Performance Under OPNET Modeler Simulation Tools in Mobile Ad Hoc Networks (MANETs). Computational Intelligence, Communication Systems and Networks, 2011 Third International Conference. ISBN: 978-1-4577-0975-3. IEEE 2011

[13] Kuinam J. Kim, Nikolai Joukov. Communications-based technology for smart grid test bed using OPNET simulations. Information Science and Applications (ICISA). Springer 2016

[14] Mingyue Ji, Giuseppe Caire, Andreas F.Molisch. Wireless Device-to-Device Caching Networks: Basic Principles and System Performance. IEEE Journal on Selected Areas in Communications - IEEE July 2015

[15] Hamnah Munir, Syed Ali Hassan, Haris Pervaiz. Energy Efficient Resource Allocation in 5G Hybrid Heterogeneous Networks: A Game Theoretic Approach. Vehicular Technology Conference - IEEE March 2017

[16] E. Bodanese. A brief introduction to heterogeneous networks (HetNets) and its challenges. Communication Technology and Application (ICCTA 2011), IET International Conference. IET- 2011

[17] Elhadj Benkhelifa, Tom Welsh, Loai Tawalbeh, Yaser Jararweh, Anas Basalamah. Energy Optimisation for Mobile Device Power Consumption: A Survey and a Unified View of Modelling for a Comprehensive Network Simulation. Mobile Networks and Applications - Springer August 2016

[18] Uchenna Odih, Panos Bakalis, Predrag Rapajic. Investigating the Impact of Traffic Type on the Performance of on Demand Routing Protocols and Power Consumption in MANET. Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015 - Springer January 2016

COMPARATIVE ANALYSIS OF CONVERGENCE TIMES BETWEEN OSPF, EIGRP, IS-IS AND BGP ROUTING PROTOCOLS IN A NETWORK

Eunice Domfeh Asabere [1], Joseph Kobina Panford [2], James Ben Hayfron-Acquah [3]

[1] Final Year M Sc. I.T, Dept. Of Computer Science, KNUST, Kumasi, Ghana, easabere@ssnit.org.gh
[2] Dept. Of Computer Science, KNUST, Kumasi, Ghana, jpanford@yahoo.com
[3] Dept. Of Computer Science, KNUST, Kumasi, Ghana, jbha@yahoo.com

---

Abstract - *Convergence time is a key factor in determining performance of routing protocols and routing protocol is one of the significant factor in determining the quality of IP communication. Convergence time is therefore very essential to a network and networks that converge faster are considered to be very reliable. The research was carried out to compare the convergence of four routing protocols namely OSPF, EIGRP, IS-IS and BGP. Network scenarios were created and a simulation was performed using Graphic Network Simulator (GNS3) to measure the convergence times of the protocols separately. Results indicated that EIGRP had the fastest convergence time in both link failure and topology change scenarios. This will help network administrators in their choice of protocols.*

***Key Words*: Convergence time, Protocol, Network, Routing, OSPF, BGP, IS-IS, EIGRP**

1. INTRODUCTION

1.1 Background of Study

Data packets traveling through the network typically traverse multiple routers and thus multiple physical links interconnecting them. Whenever there is a link failure or change in topology, routing protocols try to provide an alternative path towards the destination. It is therefore crucial that the routing protocol quickly detects such a link failure or topology change. With the increasing use of networks, any unnecessary loss of connectivity can hardly be tolerated and has to be kept as short as possible. This brings up the issue of convergence time and network is believed to have converged when the routing tables on all routers within the network are complete and correct. Routing protocols play a major role in the delivery of packets from source to destination addresses. In the study, four routing protocols namely Open Shortest Path First (OSPF), Border Gateway Protocol (BGP), Intermediate system to Intermediate system (IS-IS) and Enhanced Interior Gateway Routing Protocol (EIGRP) were compared to determine their convergence time in a given network topology.

1.2 Statement of the Problem

One of the most important characteristics of routing protocols is the convergence time The convergence time determines how fast the routers adapt their routing tables to topological changes. Among OSPF, EIGRP, IS-IS and BGP, a proof-based advice for selecting the one with the best convergence time is aimed at.

1.3 Research Objectives

The main objectives of this research are:

- To determine the convergence time for OSPF, BGP, IS-IS and EIGRP in a particular network topology.
- To compare the performance of OSPF, BGP, IS-IS and EIGRP.

1.4 Research Questions

The following are the research questions that were posed in order to accomplish the objectives.

- What is the convergence time for OSPF, BGP, IS-IS and EIGRP in a network Topologies?
- Which of the four routing protocols has the fastest convergence time in the topology used.

1.5 Significance of the Study

This study will be significant in providing an in depth understanding of the four routing protocols OSPF, BGP, IS-IS and EIGRP and determining the convergence time of these protocols in a network. The study will also compare the convergence times of these routing protocols and come out with the best one.

2. LITERATURE REVIEW

Convergence can be defined in many ways but in the context of computer networks, a network is said to have converged when all routers in a network have the same topological information about their network they find their selves in. With the help of routing protocols, routers collect topological information [1]. Convergence is a critical property in routing especially dynamic routing. There are about three forms of routing namely static, default and dynamic [2]. A network topology is said to have converged "when routing tables on all routers within the network are complete and correct" [3]. Convergence addresses the manner in which networks recover from problems and network changes. Modern networks anticipate problems by providing alternate, redundant or standby paths.

Convergence time is the time that is required for the routers in a network to learn about routes in a given network. This time is important because it helps administrators of a network to determine in the event that a network downtime occurs due to a failed link between routers or any damage to one router the amount of time it will take for that network to recover and begin to function as a normal network.

Deng et al. [4] performed analysis of RIP and OSPF and EIGRP using OPNET which is a simulator widely used for networking related analysis. In their research, they analysed the performance of these protocols based on their convergence activity, convergence duration and traffic sent

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(bytes/sec) to compare the difference in their performance. From their research, they found out that the convergence of EIGRP was faster than the others regardless of the network topology.

Panford *et al*. [5] also analyzed Convergence times between RIP and EIGRP routing protocols in a network using packet tracer which allows network behaviour experimentation and also helps in answering what-if scenarios. In their research, they observed that EIGRP had the fastest convergence time. Their experiments also showed that, regardless of the topology, the convergence time remains the same whether for RIP or EIGRP. Another interesting observation made with EIGRP was that as the number of routers increases, the time for convergence were almost the same.

3. METHODOLOGY

The method for this research was a simulation of scenarios. To help with this simulation, Graphical Network Simulator (GNS3) was employed and the network diagram for the simulation scenarios is illustrated in Fig.1. GNS3 was chosen because it has a user- friendly Graphical User Interface (GUI) and also enables users to configure a network component in a virtual machine that runs the OS same as the original network component.

4. ANALYSIS

The measurements results were placed into three main categories. The first category, based on Fig. 2 consist of measurements of convergence times of protocols with link failure closer to the source of the traffic as shown in Table 1. The second category as derived from Fig 3. is made up of measurements of convergence times of protocols with link failure closer to the destination of the traffic as illustrated in Table 2 and the last category was convergence time measurements under topology change as shown in Table 3. Fig. 4 and Fig. 5 show the network diagram as additional routers are added to the original network diagram.

4.1 Results of Routing protocols with Failure closer to the source of the traffic

Table -1: Convergence time measurement for Protocols with Link Failure closer to the source of the traffic.

Test	OSPF	EIGRP	IS-IS	BGP
1	8.346	6.078	7.799	21.975
2	8.637	5.985	8.340	23.169
3	8.494	5.938	8.368	19.331
4	8.879	5.951	8.185	15.275
5	8.162	5.469	8.590	15.743
6	9.601	6.039	9.095	19.874
7	7.788	5.491	8.532	25.646
8	8.592	5.169	8.042	28.507
9	8.836	7.298	8.051	14.664
10	6.187	6.204	7.441	28.640

4.2 Results of Routing protocols with Failure closer to the destination of the traffic

Table -2: Convergence time measurement for Protocols with Link Failure closer to the source of the traffic.

Test	OSPF	EIGRP	IS-IS	BGP
1	7.993	4.220	8.091	13.553
2	8.490	3.969	8.485	29.517
3	7.797	4.079	8.344	31.220
4	9.078	4.007	8.438	29.993
5	9.005	5.968	8.905	16.394
6	8.673	5.972	8.297	18.859
7	8.938	4.298	7.801	27.502
8	8.841	6.001	8.438	20.646
9	7.735	5.875	8.660	17.641
10	7.704	5.969	7.438	24.204

4.3 Results of Topology Change

Table -3: Convergence time measurement for Protocols with Topology Change

	OSPF	EIGRP	IS-IS	BGP
1	12.138	3.027	8.044	18.227
2	11.117	3.031	12.148	20.258
3	10.913	3.468	11.082	18.229
4	11.530	3.477	9.275	19.541
5	10.026	3.102	9.196	19.205
6	10.993	3.198	10.084	18.714
7	11.362	3.144	9.143	21.095
8	11.122	3.112	8.012	19.521
9	11.212	3.099	8.050	18.008
10	10.410	3.005	7.048	18.035

On the average, it took OSPF network 8.352s to converge, EIGRP network 5.962s to converge, IS-IS network 8.244s to converge and BGP network 21.282s to converge for link failure closer to source of the traffic. For link failure close to the destination of traffic the average convergence times were 8.375s for OSPF, 5.036s for EIGRP, 8.290s for IS-IS and 22.953ms for BGP.

It took an average time of 11.082s for OSPF to converge, 3.166s for EIGRP to converge, 9.208s for IS-IS to converge and 19.083s for BGP to converge for change in topology.

5. CONCLUSION AND RECOMMNDATION

From the simulation results the EIGRP give the best performance. EIGRP generate the least traffic and thus it will consume the least bandwidth, leaving enough bandwidth for transmission of data. EIGRP also has the best performance in the case of topology changes and when there is a broken Ethernet connection.

In conclusion, the simulations confirmed that EIGRP was the best choice for all scenarios implemented as it has a fast convergence, while also efficiently utilizing bandwidth.

IS-IS was the second choice as far as convergence time was concerned and then OSFP came next. BGP performed poorly

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and is therefore not suitable for large networks. It can therefore be stated based on the results achieved that there is a significant difference in the performance of the protocols as far as convergence time is concerned.

REFERENCES

[1] Shah A, Waqas J. Rana, "Performance Analysis of RIP and OSPF in Network Using OPNET", International Journal of Computer Science Issues, Issue 6, No 2, November 2013.

[2] Lammle, Todd, "CCNA Cisco Certified Network Associate study guide, sixth edition". Indianapolis, Ind.: Wiley. (2007).

[3] Todorovic Ivana, Sepanovic Stevan, "Measurements of convergence time for RIP And EIGRP Protocols", Scripta Scientiarum Naturalium, volume 2, 2011.

[4] Deng Justice, Wu Siheng, Sun Kenny, "Comparison of RIP, OSPF and EIGRP Routing Protocols based on OPNET" Final project. Spring, 2014.

[5] Panford, J. K., Riverson K. & Boansi O. K (2015). Comparative analysis of convergence times between rip, and eigrp routing protocols in a network. Research Journali's; Journal of Computer Science., pg 1-5.

APPENDIX

The following are network topologies used in the experiments.

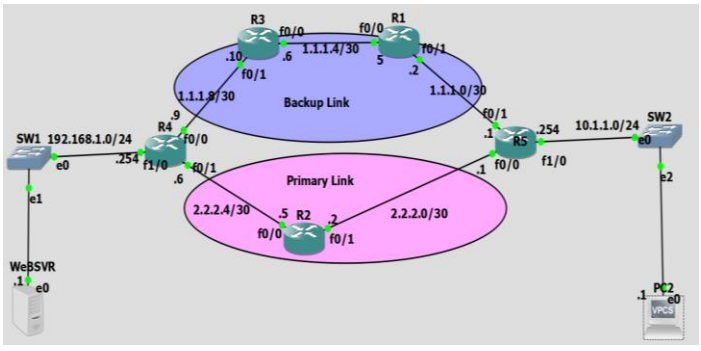


Fig -1: Network diagram for simulation scenarios

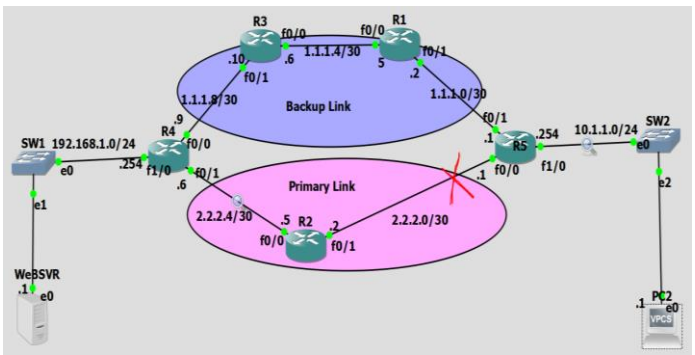


Fig -2: Network diagram for link failure closer to the source of traffic

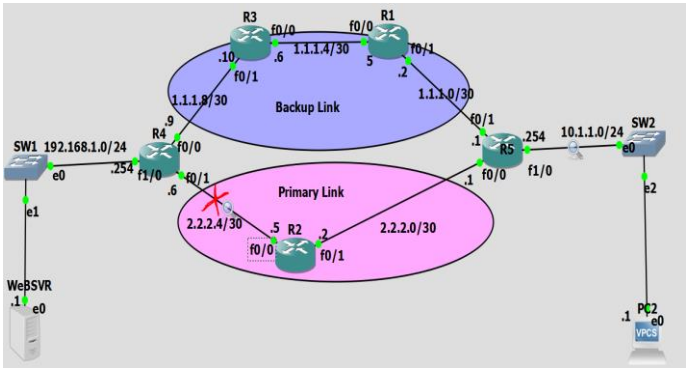


Fig -3: Network diagram for link failure closer to the destination of traffic

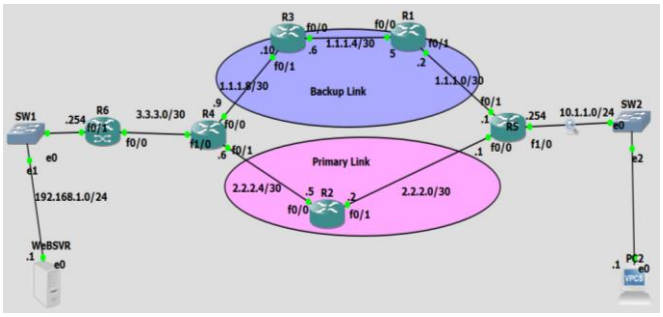


Fig -4: Network diagram with one additional router

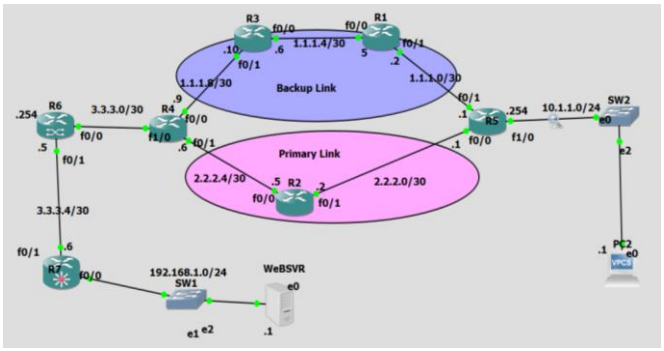


Fig -5: Network Diagram with two additional routers

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ARABIC SPELLING CHECKER ALGORITHM FOR SPEECH RECOGNITION

Huda F. Al-shahad

Department of Computer Science
University of Kerbala
Karbala, Iraq
huda_msc2006@yahoo.com

ABSTRAT

The Automatic Speech Recognition is defining as the process of convert a speech wave into text by using a computer. Speech recognition is the easiest way manipulate with the computer application especially to the people that have no arms. This paper proposes an Arabic word and popular language (Iraqi language) error correction method and algorithm for speech recognition system. The proposed algorithm is split the input content (that is input as a speech wave and convert it to text by speech recognition system) into a few word-tokens that are submitted as search questions to the system. The system offer to replace the error word by the suggested correction using n-gram features and save the writing words in a text file that the user will choose the path of it. Future research can improve upon the proposed system so much so that it can be take many correction algorithms and make difference between them.

Keywords: *Speech Recognition; Arabic Error Correction; popular language; Token; n-gram.*

I. INTRODUCTION

Speech technology is presently broadly utilized as a part of the field of discourse chronicling, for example, PodCastle [1]. In these frameworks, the words are perused by client or to recover the fitting sections utilizing watchwords, a low word-blunder rate (WER) is hardly must require, so the model must the most suitable words between the hopefuls assumed by an automatic speech model. Be that as it may, if many words in model are false, it might be chosen independent of what is the dialect display. This problem need solve, a few distinguish language models [3, 4, 5] have been proposed to re-rank the N-best sentences after large-vocabulary, continuous speech recognition.

The use of N-grams trained from speech recognition results including false words and it given transcription. This paper describes a method that receive the text from speech system after convert it and correct the error words by suggest the correction words that make the user have the flexibility to choose any one or replace the error word with the true word. After that the correction words will save in a text file. Many propelled discourse acknowledgment frameworks utilize trainable dialect models that can be advanced for a specific (speaker-free) and in addition for a particular sub-dialect use. This enhancement is important to accomplish a respectable level of acknowledgment precision; be that as it may, it may not ensure reliably high-exactness execution because of the constrained abilities of the basic dialect display, generally 2-or 3-gram HMMs. The method in this paper is to take the Arabic and popular language (Iraqi language) word (one or more) that the speaker says and make the correctness on it if it error. Unlike other approaches (e.g., Bassil& Alwani, 2012) that use the suggestion technology Bing's spelling to recognize error and correct the words that input by automatic speech recognition that recognized output text[5]. The other paper (Nishizaki & Sekiguchi, 2006) describes an error correction method of continuous speech recognition using WEB documents for spoken documents indexing [6]. Fusayasu, i Tanaka, Takiguchi and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Ariki in 2015 focus on their research on a word-error correction system for continuous speech recognition using confusion networks[7].
The structure of this paper is as follows. In Section 2, discuss Arabic challenges. In Sections 3 the error detection of text, the methodology is discussed in section 4, error correction is described in Section 5 and in 6 Computation algorithm is describe, and the experimental results are shown in section 7. Finally, the conclusion is view in Section 8.

II. CHALLENGES OF ARABIC SPEECH

Arabic speech recognition faces many challenges. One of these is the vowels of Arabic word are short which are may be ignore in text. Another one is Arabic language has many tones where each word is pronounced in a different way.
Arabic many-sided quality is appearing by the expansive number of affixes (prefixes, infixes, and suffixes) that can be added to the three shape design. Farghaly and Shaalan in 2009 gave an investigation of Arabic dialect difficulties and answers for it[8]. Lamel et al. in 2009 introduced many number of difficulties for Arabic discourse acknowledgment, for example, very large lexical variety[9].

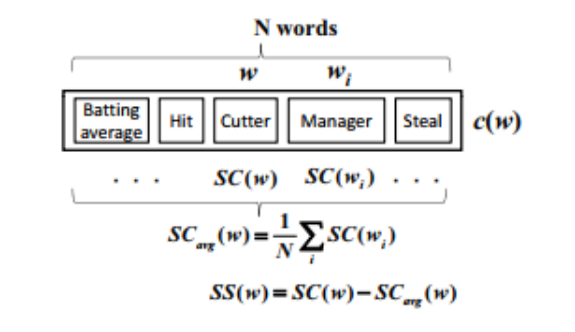


Figure (2) Semantic score

Similarity SC(wi) between the context c(w) and the number of word wi in the context is computed by latent semantic analysis (LSA) [10].

III. LATENT SEMANTIC ANALYSIS

" Latent Semantic Analysis (LSA) is a hypothesis and technique for separating and speaking to the importance of words. Significance is evaluated utilizing factual calculations connected to a huge corpus of content. The corpus encapsulates an arrangement of common limitations that to a great extent decide the semantic likeness of words and sets of words. These requirements can be understood utilizing direct variable based math strategies, specifically, Singular Value Decomposition."4 LSA is a numerical and factual approach, guaranteeing that semantic data can be gotten from a word-record co-event network and words and reports can be spoken to as focuses in a (high-dimensional) Euclidean space.

Dimensionality diminishment is a basic piece of this inference. LSA depends on the Vector Space Model (VSM), a mathematical portrayal of content archives generally utilized as a part of data recovery. The vector space of an accumulation of writings is built by speaking to each report as a vector containing the frequencies of the words or terms the record is made out of as components. By and large, these archive vectors signify a term-by-report framework speaking to the full content accumulation. Relatedness of records can be gotten from those vectors, e.g. by figuring the edge between archive vectors by methods for a cosine measure. In any case, this numerical portrayal of content information does not illuminate commonplace issues of working with dialect. From one perspective there are morphological issues for the correct recognizable proof of terms and the way that not all terms in content are of equivalent significance. This can be settled by highlight determination methods (stemming, stop word expulsion, collocations, equivalent word records, space vocabulary, grammatical form taggers, and data pick up) and weighting plans (TF-IDF, Log-Entropy). Solitary Value Decomposition (SVD) is utilized as a rank bringing strategy down to truncate the first vector space to uncover the hidden or 'inactive' semantic structure in the example

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

of word utilization to characterize archives in a gathering. This truncation permits managing common dialect issues like synonymy as various words communicating a similar thought should be near each other in the diminished k-dimensional vector space. SVD will break down the first term-by-report network into orthogonal components that speak to the two terms and records:

$$\mathbf{A = U \Sigma VT} \qquad \mathbf{(1)}$$

With A the original term-by-document matrix, Σ a diagonal matrix with the square roots of singular values of A.AT and AT .A ($\sigma 1^2 > \sigma 2^2 > \ldots > \sigma n^2$), and U and V containing left and right singular vectors [].

We will generate the document-word matrix by using tf-idf as shown in the following equation:

$$\mathbf{TFIDFi,j = (Ni,j / N^*,j) * log(D / Di)} \quad \mathbf{(2)}$$

After that the document is factored using singular value decomposition (SVD) as follows;

$$\mathbf{W = USV^T} \qquad \mathbf{(3)}$$

Using the row vector ui of the matrix U and the row vector vj of the matrix V , the similarity sim(ri, cj) between the document cj and the word ri is computed as follows:

$$\mathbf{sim(ri, cj)} = \frac{\mathbf{uiSvT\ j}}{|\boldsymbol{uiS}^{\frac{1}{2}}\ ||\boldsymbol{uiS}^{\frac{1}{2}}\ |} \qquad \mathbf{(4)}$$

IV. COMPUTATION ALGORITHM

The semantic aim of any word is defined to be high if the meaning of the selected word is similar to the meaning of the words around the underlining word. The semantic result of the word w is computed as follows:

(1) c(w) is represent the context of content word w that framed as the gathering of the substance words around w including a similar word, as appeared in Figure(2).

(2) The similarity SC(wi) between the context c(w) and the word wi in the context is computed, where i represent the number of word.

(3) The average similarity SC(wi) is computed as SCavg(w).

(4) Normalized similarity SS(w) is computed the difference between SC(w) and SCavg(w) as shown in equation below;

$$SS(w) = SC(w) - SCavg(w)$$

V. ERRORS DETECTION

Error detection problem may be solved by two techniques :*N-gram* analysis and *dictionary lookup*. Error correction method consists of checking to know if the input string is valid or not.

In this paper the n–gram will use. N-gram method is characterized as a strategy to discover erroneous words in content. Rather than looking at each time every whole word in content to the appropriate lexicon, n-grams will control the checking. The checking is finished by utilizing a matrix with an n-dimensional size where the frequencies of real n-gram are put away. On the off chance that a nonexistent or unusual n-gram is discovered then the word is flagged as an incorrect spelling word, else not.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A n-gram is an arrangement of successive characters brought from a content with a length of n as is set to.
When n = 1 character then the term that utilized is a **Unigram**,
When n = 2 characters then the term that utilized is a **Bigram**,
When n = 3 characters then the term that utilized is **Trigram**.

VI. MRTHODOLOGY

The proposed method in this paper shows the process of the error detection model using N-gram and LSA information to show the cost of error correction in system as shown in figure (1). The first step, speech data are recognized and the recognition results are output as a token. Second, each word is marked as false or true. After the recognition errors, the system will suggest the correction words that make the user choose which correct one need, and then the words will save in a text file in any drive preferred.
In this paper, as said above, word-mistake amendment can be accomplished in the perplexity set by choosing the word with the most astounding estimation of the accompanying straight discriminant work. We utilize the best probability words in the disarray arrange if the perplexity set has no third probability word, it is supplemented with the second one. Also, on the off chance that it has no second probability word, it is supplemented with the first. After the learning procedure is done, acknowledgment blunders are rectified utilizing the calculation beneath:

(1) Receive the text as voice, so we will convert the voice to test.

(2) Make tokenize and concentrate the best probability words from the perplexity organize and detect recognition.

(3) Using the error detection model, “N-gram”.

(4) Apply the LSA algorithm to find the suitable word.

(5) Select the best likelihood word in the confusion set if the word identified as correct data does not exist.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

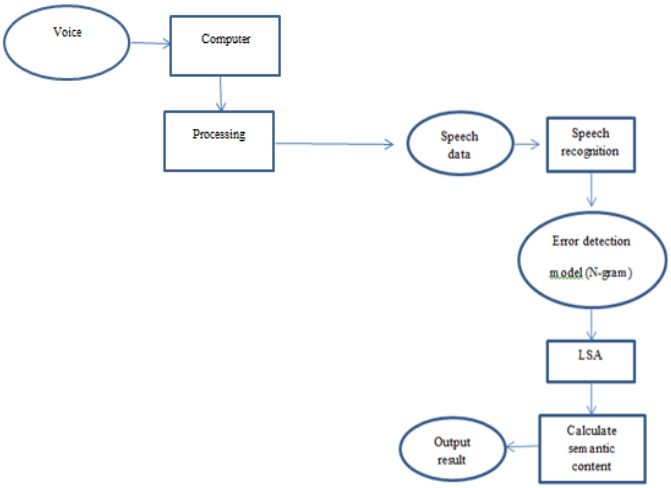


Figure (1) proposed method

VII. ERROR CORRECTION

The proposed error correction algorithm includes several steps that must execute in order to detect the error and then correct it. The algorithm take-off by divide the recognized output transcript into many tokens T={ t1…tn }, each composed of n words, ti={w0,w1,w2,w3,w4,..wn} where ti is represent the special token and wj is a single word in that token. Then, every ti is sent to check the validation of it using n-gram to check the ranking and show the suggestion correct words ci. If the word is valid, at that point token ti must not contain a specific incorrectly spelled word; and thus, ti is supplanted by ci. At last, after all tokens get approved, all the first right tokens O={ t1…tk }, plus the corrected ones C={ c1…cp } will concatenate with each other, to make a new text with fewer error represented formally as V={ v1…vk+p }. In this paper, we use the characteristic N-gram. To put it plainly, we utilize it to distinguish acknowledgment blunders. This sort of discriminative dialect demonstrate can be prepared by fusing the discourse acknowledgment comes about what's more, the comparing right interpretation. Discriminative dialect models, for example can distinguish unnatural N-grams and adjust the false word to fit the characteristic N-gram.

VIII. EXPERIMENT AND RESULTS

In the experiments, speech recognition was performed in to two different languages: English and Arabic (clear Arabic language and Iraq language).The proposed calculation was executed utilizing MS C# 4.0 under the MS .NET Framework 4.0 and the MS Visual Studio2012.The speech recognition will enter the words that the user want to print as in figure (3) and then the system will lexical the text into many tokens, take the token to check if the word error or not, if the word is error many suggestion will appear to replace it by the word as in figure (4). The correct output will save in word text, the figure(5) shows the Iraqi word that the user may be entered when talk with family in another place for example or when two friends talk to each other by computer or through any social media program.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

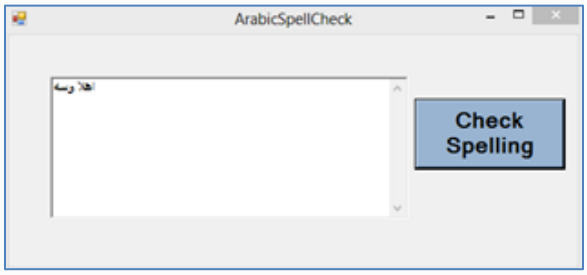


Figure (3) Speech Input

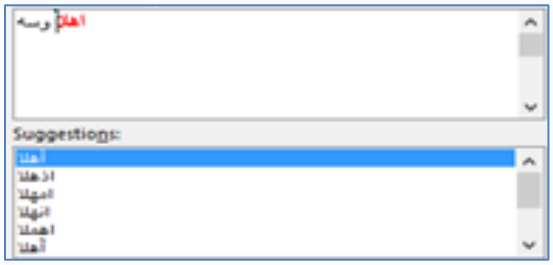
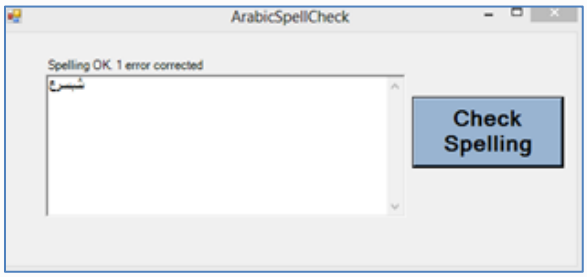
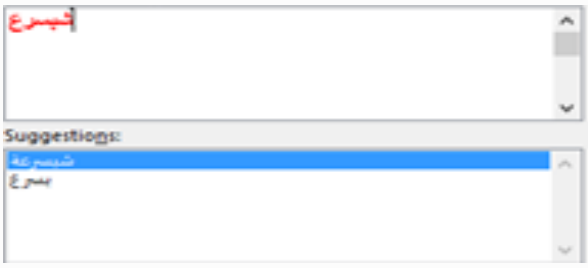


Figure (4) Correction suggestion



figure(5-A) Iraqi words



figure(5-B) Iraqi word correction

When we use the LSA algorithm the count matrix that output from the Arabic text is as follow:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
[[ 1.  2.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 2.  1.  0.  0.  1.  0.  0.  0.  0.]
 [ 0.  1.  1.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 0.  1.  0.  0.  2.  2.  2.  0.  2.]
 [ 0.  0.  0.  0.  0.  1.  1.  0.  1.]
 [ 0.  0.  0.  0.  0.  1.  1.  1.  1.]]
```

figure(6) count matrix

The reason SVD is valuable, is that it finds a diminished dimensional portrayal of our lattice that underscores the most grounded connections and discards the commotion. As it were, it makes the most ideal recreation of the framework with the minimum conceivable data. To do this, it tosses out commotion, which does not help, and underlines solid examples and patterns, which do help as shown in follow:

```
Here are the singular values
[  5.47054849e+00   3.24054292e+00   1.54975161e+00
   8.69865786e-01   7.43977135e-01   6.15341902e-17
   4.27433567e-35]
```

figure(7) SVD vector

After that the similarity will use to check the validity of the word in the document, the figure (8-a) show the part of columns of the matrix and figure (8,b) show the part of raws of the matrix.

```
[[ 0.07868054 -0.61603335  0.50014023]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.1064358  -0.6735659  -0.46037367]
 [ 0.03566079 -0.23712335  0.57850259]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.73209861 -0.11018105 -0.30750493]
 [ 0.29633239  0.1373077   0.1251845 ]
 [ 0.30657657  0.15175944  0.21449164]]
```

figure(8-a) column of matrix

```
[[  5.32948634e-02   1.88565420e-01   6.51868729e-03  -2.77555756e-17
    2.87107044e-01   5.40366902e-01   5.40366902e-01   5.60412855e-02
    5.40366902e-01]
 [ -6.05813657e-01  -6.95234426e-01  -7.31739566e-02   1.38777878e-17
   -2.75857479e-01   1.48317161e-01   1.48317161e-01   4.68314850e-02
    1.48317161e-01]
 [ -2.71402911e-01   5.23248020e-01   3.73287298e-01  -0.00000000e+00
   -6.93907026e-01   6.46682950e-02   6.46682950e-02   1.38403886e-01
    6.46682950e-02]]
```

figure(8-b) raws of matrix

IX. CONCLUSION

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In this paper, I have proposed an automatic speech recognition Arabic and Iraqi language error correction by using n-gram algorithm. The proposed two-step first, speech data are recognized and split the text into tokens , second, each word is labeled as false or true and recognition errors, the system will suggest the correction words, correction method can efficiently use the n-gram method.

REFERENCES

[1] M. Goto, J. Ogata, and K. Eto, "Podcastle: A web 2.0 approach to speech recognition research," in Proc. Interspeech2007, 2007, pp. 2397–2400.

[2] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in Proc. ACL, 2004, pp. 47–54.

[3] T. Oba, T. Hori, and A. Nakamura, "A study of efficient discriminative word sequences for reranking of recognition results based on n-gram counts," in Proc. Interspeech2007, 2007, pp. 1753– 1756.

[4] A. Kobayashi, T. Oku, S. Homma, S. Sato, T. Imai, and T. Takagi, "Discriminative rescoring based on minimization of word errors for transcribing broadcast news," in Proc. ISCA, 2008, pp. 1574– 1577.

[5] Y. Bassil and M.Alwani, "Podcastle: Post-Editing Error Correction Algorithm For Speech Recognition using Bing Spelling Suggestion," in (IJACSA) International Journal of Advanced Computer Science and Applications2012, pp.95-101.

[6] H. Nishizaki and Y. Sekiguchi, "Word Error Correction of Continuous Speech Recognition Using WEB Documents for Spoken Document Indexing"2006, ICCPOL pp. 213–221.

[7] Y. Fusayasu, K.Tanaka, T. Takiguchi and Y. Ariki," Word-Error Correction of Continuous Speech Recognition Based on Normalized Relevance Distance",in Conference on Artificial Intelligence (IJCAI 2015), pp.1257-1262.

[8] A.Farghaly and Kh.Shaalan," Arabic Natural Language Processing: Challenges and Solutions ", 2009, ACM Transactions on Asian Language Information Processing, Vol. 8, No. 4.

[9] L. Lamel, A. Messaoudi et al ," Automatic speech-to-text transcription in Arabic ",2009, ACM Transactions on Computational Logic, pp. 1–17

[10] T. Landauer, P. W. Foltz, D. Laham, "Introduction to Latent Semantic Analysis", in Discourse Processes, 1988, pp. 259–284.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Detection of physiological changes in women during muscle activity using electromyography and other techniques

Rossana Rodríguez Montero[a], David Asael Gutiérrez Hernández[a], Miguel Mora González[b], Víctor Manuel Zamudio Rodríguez[a], Juan Martín Carpio Valadez[a], Manuel Ornelas Rodríguez[a], Claudia Margarita Lara Rendón[a], Miguel Salvador Gómez Díaz[a], Jesús Ricardo Sevilla Escoboza[b], and Víctor Porfirio Vera Ávila[b]

a. Tecnológico Nacional de México. Instituto Tecnológico de León. León, Guanajuato, México.
b. Universidad de Guadalajara. Centro Universitario de los Lagos. Lagos de Moreno, Jalisco, México.

***Abstract*— Bio-signals are important to know what is going on with our body. Especially the muscular activity is related with physiological changes in a woman, for example with their menstrual cycle. Besides of this, it is required to evaluate muscular activity over this changes, this can be done with electromyography (EMG) and the entropy, which allow the comparison of the obtained signals to measure those physiological changes.**

In previous works, muscular fatigue has been evaluated with EMG; nevertheless it has not been going into deep with chemical changes that are produced into the body, in a natural way which can alter the obtained bio signal coming from the muscle.

We developed a digital portable electromyograph to get electromyography samples. By means of it, the women´s bio-signals were studied, for those who were under an exercise routine and also for those who were not.

While visualizing the behavior of the electromyography obtained from the muscle, we perceived the singularity that the bio-signal for women of both group, while being on a menstrual cycle were similar.

Thus, it was implemented the entropy on the signals to justify the results obtained on the electromyography and the personal test applied. As a result, we proved that these signals are really showing one of the physiological changes in a woman.

***Index Terms*— Electromyography (EMG); Entropy; Membrane potential; Muscle activity.**

I. INTRODUCTION

In the human body, the biological signals can come from many physical phenomes. For the case of this work, the bio-signal is taken from the biceps muscle area; but to be able to process and obtain conclusions from the muscle activity, first it has to be converted in signals of electric character [1]. To achieve the obtainment of that signal, electromyography is the selected technique to be applied.

The EMG is a biomedical signal that measures the electric current generated in the muscles during its contraction and represents the muscular activity. One of the most popular techniques for the acquisition of this kind of signals is the superficial electromyography, that is commonly used for many researches, being this a noninvasive technique that uses electrodes that are put on the skin environment for taking the differential of bio-potential created by the variations of current in the muscle cells [2,3] which can be useful for quantitative technique for evaluation and registration of the electric activity produced by muscles [1,4].The electromyography signals contains relevant information that may be used for patron detection of a signal [5], the progress of the muscular fatigue [6,7], among other systems or applications type[3].

We can say that this signals give a time serial of biological kind; on which the time in a signal point, can mean more than a

This work was supported in part by the Conacyt, and the Instituto Tecnologico de Leon's laboratory "LACIT".

Rossana Rodríguez Montero is with the Instituto Tecnológico de León, León, Guanajuato, México (e-mail: m105060381@itleon.edu.mx).

David Asael Gutierrez Hernandez, is with Department of postgrade and Intelligent environments department, Instituto Tecnológico de León, León, Guanajuato, Mexico (e-mail: david.gutierrez@itleon.edu.mx).

Miguel Mora González is with Department of Exact Sciences and Technology, Centro Universitario de los Lagos, Universidad de Guadalajara. Lagos de Moreno, Jalisco, México. (e-mail: mmora@culagos.udg.mx).

Víctor Manuel Zamudio Rodríguez is with Intelligent environments department, Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: vic.zamudio@ieee.org).

Juan Martín Carpio Valadez is with Intelligent Systems Department, Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: juanmartin.carpio@itleon.edu.mx).

Manuel Ornelas Rodríguez is with Intelligent Systems Department, Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: manuel.ornelas@itleon.edu.mx)

Claudia Margarita Lara Rendón is with Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: claus.l.r@icloud.com)

Miguel Salvador Gómez Díaz is with Instituto Tecnológico de León, León, Guanajuato, Mexico. (e-mail: miguel_gomez@itleon.edu.mx)

Jesús Ricardo Sevilla Escoboza is with Department of Exact Sciences and Technology, Centro Universitario de los Lagos, Universidad de Guadalajara. Lagos de Moreno, Jalisco, México. (e-mail: jesus.sescoboza@academicos.udg.mx).

Victor Porfirio Vera Avila is with Centro Universitario de los Lagos, Universidad de Guadalajara. Lagos de Moreno, Jalisco, México. (e-mail: viktorvera.182@gmail.com).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

simple analysis [8] in the muscular activity made.

The biological temporary serial data not necessarily can be evaluated with common methods used from time series analysis, like the techniques of autocorrelation and frequency domain [9].

With the previous explanation, the objective of this work is to identify changes or physiological states through the computational exploration of biological signals, obtained during muscular activity and by means of electromyography and entropy technique to evaluate this signal.

II. Theoretical Description

For the development of this work some theoretical concepts must be considered. There are the following:

A. *The membrane potential*

Neurons are the basic functional units of the nervous system, and they generate electrical signals called action potentials, which allow them to quickly transmit information over long distances.

The different classes of neurons that found in the human nervous system, can be divided into three classes: sensory neurons, motor neurons, and interneurons; where these have three basic functions, these are to receive signals (or information), integrate incoming signals (to determine whether or not the information should be passed along) and communicate signals to target cells (other neurons or muscles or glands).

That conjunction of neurons, when your brain decides to move a muscle, motor cortex neurons travel through the spinal cord to synapse with "lower motor neurons." These motor neurons on moment to make synapse with the muscle form a "motor unit", where a motor unit is composed of an individual motor neuron and many muscle fibers it innervates. A muscle fiber is a very special cell type that can change its shape thanks to the actin / myosin chains that travel in it [10].

An individual motor neuron can synapse with many muscle fibers. In general, a large muscle such as the biceps has motor neurons that innervate thousands of muscle fibers while other muscles, such as those in the eye, which require a lot of precision, have motor neurons that innervate less than ten muscle fibers [11].

When a motor neuron triggers an action potential, this potential generates a release of acetylcholine (fig. 1-2) at the synapse between the neuron and the muscle (this synapse is also known as Neuromuscular Junction). Acetylcholine causes a change in the electrical potential of the muscle. When this electric potential reaches a threshold, an action potential is generated in the muscle fiber this action potential propagates through the muscle membrane, causing the voltage-dependent calcium channels to open, which begins the cellular cascade that finally generates muscle contraction.

When you contract a muscle, it is because many muscle fibers are firing action potentials and changing their shape (fig. 2).

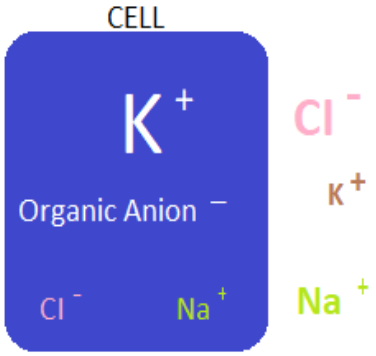


Fig. 1. Diagram of concentration and change of anions and cations, sodium (Na^+), potassium (K^+), chloride (Cl^-). (Image modified from "The sodium-potassium exchange pump," by Blausen staff (CC BY 3.0)).

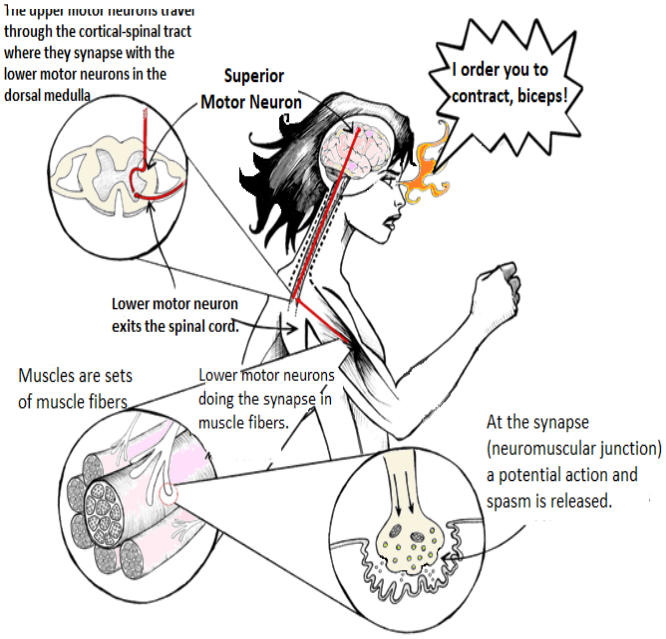


Fig. 2. Diagram of electrochemical process for the generation of muscle movement. Own elaboration

B. *Entropy*

According to [12] entropy is the degree of disorder that a system has and can be considered as a measurement standard. Entropy can be considered as a measure of uncertainty, so that the information needed in any process can be narrowed, reduced or eliminated uncertainty.

Entropy generation clarifies energy losses in a system evidently in many energy-related applications. Bejan [13] originally formulated the analysis of entropy generation.

1) *Permutation Entropy*

Permutation Entropy (PE) was introduced as a complexity parameter for time series based on comparison of neighboring values; the advantages are its simplicity, extremely fast calculation and robustness [14]. That kind of entropy is an appropriate complexity measure for chaotic time series, in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

particular in the presence of dynamical and observational noise. In contrast with all known complexity parameters, a small noise does not essentially change the complexity of a chaotic signal. Permutation entropies can be calculated for arbitrary real–world time series. As the article [15] says.

The article [16] says, that the algorithm to compute the PE can be divided into four basic steps, as the article:

1. Fragment the continuous EEG signal into segments containing *m* samples (*m* is called the embedding dimension); for a given embedding dimension *m* = 3 there will be *m*! possible permutations called motifs, so in this case six different motifs are obtained.
2. Identify each motif as belonging to one of the six different categories.
3. Obtain the probability of occurrence of each motif in the signal (pi) by counting the number of motifs of each of the six different categories.
4. Apply the standard Shannon uncertainty formula to calculate the PE of the resultant normalized probability distribution of the motifs (Eq. 1).

$$PE = \frac{-\sum(p_i \cdot \ln(p_i))}{\ln(number\_of\_motifs)} \quad (1)$$

C. Fourier transform

The Fourier transform (FT) has been approached from the formulation of the discrete signal, closer to its use in computable methods and algorithms, with its practical side of tool creation and applications in the treated field. The Fourier transform [17] represents a useful tool to extract the information contained in a signal on the domain frequency. FT is provided by its integral [18], that this provides a frequency function. That function is complex: its module is the spectral amplitude and the square from the amplitude is the density of spectral power (DSP) [19]. This spectral density is, the Fourier transform of the autocorrelation. The spectra term is used for the amplitude and for the power density represented in front of the frequency [19].

The DSP can be determined by some methods. The most used are the entropy method and the square of the Fourier transform. That is the reason why we used this two methods.

D. Hilbert Huang transform

The article [20] says, that Hilbert-Huang transform (HHT) is NASA's designated name for the combination of the empirical mode decomposition (EMD) and the Hilbert spectral analysis (HSA). It is an adaptive data analysis method designed specifically for analyzing data from nonlinear and nonstationary processes. The key part of the HHT is the EMD method with which any complicated data set can be decomposed into a finite and often small number of components, called intrinsic mode functions (IMF).

HHT is an empirical approach, and has been tested and validated exhaustively but only empirically. In almost all the cases studied, HHT gives results much sharper than any of the traditional analysis methods in time-frequency-energy representation [20]. Additionally, it reveals true physical meanings in many of the data examined. The Hilbert–Huang Transform (HHT) is a new time–frequency analysis method [21]. The main difference between the HHT and all other methods is that the elementary wavelet function is derived from the signal itself and is adaptive. The main feature of the HHT is the Empirical Mode Decomposition (EMD), which is capable of extracting all the oscillatory modes present in a signal. Each extracted mode is referred to as an Intrinsic Mode Function (IMF), which has a unique local characteristic [22, 23]. After the Hilbert transform on each IMF has been performed, the time–frequency distribution of the signal energy is obtained, which is referred to as the Hilbert spectrum.

III. Experimental Setup

For this experiment we used a digital electromyograph, developed in the laboratory of LACIT of the Technological Institute of Leon that serves as our system of acquisition of the muscular signal. This instrument has three cables that receive the signal, where an electrode is connected to each cable, these electrodes are placed on the skin of the test subject (figure 3); where two of them are in the middle and lower biceps (toward the elbow, calling these as positive and negative) and the third electrode is placed in the elbow area, as shown in figure 4; the reading obtained is saved in an extension file ".txt", then the information is analyzed on a computer. All of the above is the acquisition technique; for this work four samples were taken, one per week; seven women participate in the experiment, where they separate into 2 groups, those who exercise and those who do not exercise, their average age is 24.8 years. They responded to a brief questionnaire with their physical characteristics, performed exercise, and in the physiological phase that was (in menstrual cycle or not), which we suppose could influence the muscular response, by the chemical

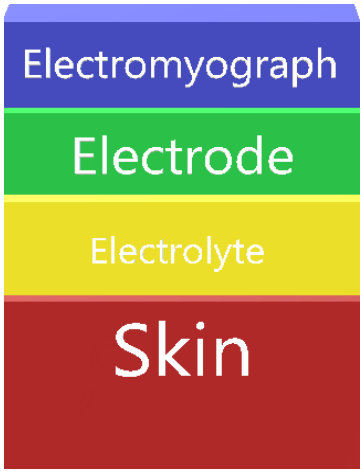


Fig. 3. Diagram of the Electromyograph connection. Own elaboration

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

elements that are in the process of communication as illustrated in figure 2. Their responses are shown in Table 1.

The experiment consists in performing weightlifting, of 3 sets of 12 repetitions with time intervals of 30 bits per second, where the first 10 seconds are left in basal mode (resting muscle), after which the first series of 12 repetitions, in the second 56 approximately the rest is performed, and the second series starts at approximately 01:12 minutes, rests and the third series starts at approximately 02:12 minutes. The samples taken were 4 for each woman, one sample for each week. The weight's dumbbells is seven pounds. The full time of the experiment is about 3:15 minutes.

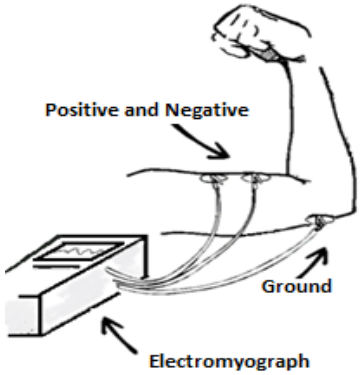


Fig. 4. Configuration to take electromyographic signals. Own elaboration

TABLE I
PEOPLE DESCRIPTION FOR THIS EXPERIMENT

Features	Woman 1	Woman 2	Woman 3	Woman 4	Woman 5	Woman 6	Woman 7
Age (year old)	22	33	26	21	23	24	25
Weight (kg)	60	68	63	67	94	89	69
Height (m)	1.73	1.62	1.60	1.66	1.68	1.73	1.67
Are you on period? (week 1)	Yes	No	No	No	No	No	No
Are you on period? (week 2)	No	No	Yes	No	No	No	Yes
Are you on period? (week 3)	No	Yes	No	No	No	Yes	No
Are you on period? (week 4)	No	No	No	Yes	Yes	No	No
Do you practice exercise?	Yes	Yes	Yes	No	No	No	No
Which sport do you practice?	Swimming	Spinning	Dumbbells and box	N/A	N/A	N/A	N/A
How many days do you exercise on a week?	2	5	5	N/A	N/A	N/A	N/A

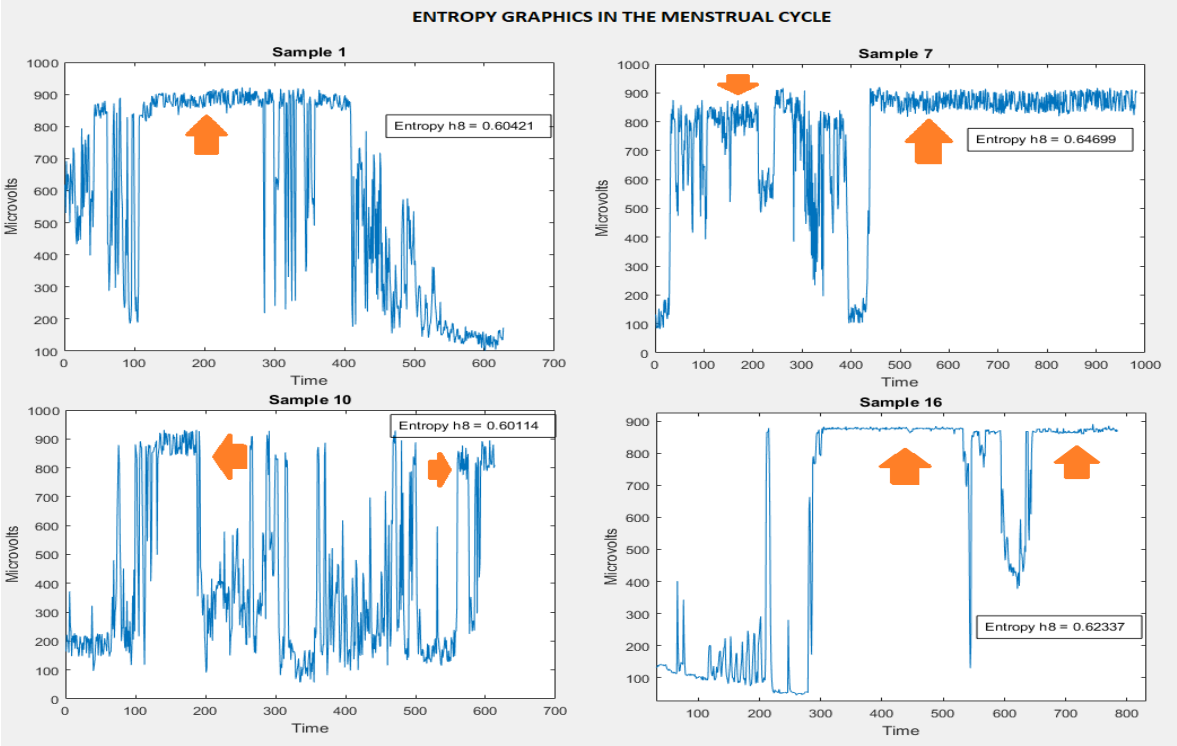


Fig. 5. Some graphics that shows the women period, and evaluation with entropy being 8 points of permutation.

Once the signal from each experiment is obtained, it was inferred that the physiological phase (woman's period) can be recognized in these signals (figure 5). This based on the information shown in table 1; the description of the chemicals that travel through the body and allow movement (membrane potential); the same signaling of it, the areas with orange arrows (figure 5) that mark the variations of the signal when the women were found in their period, unlike when they were not (figure

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

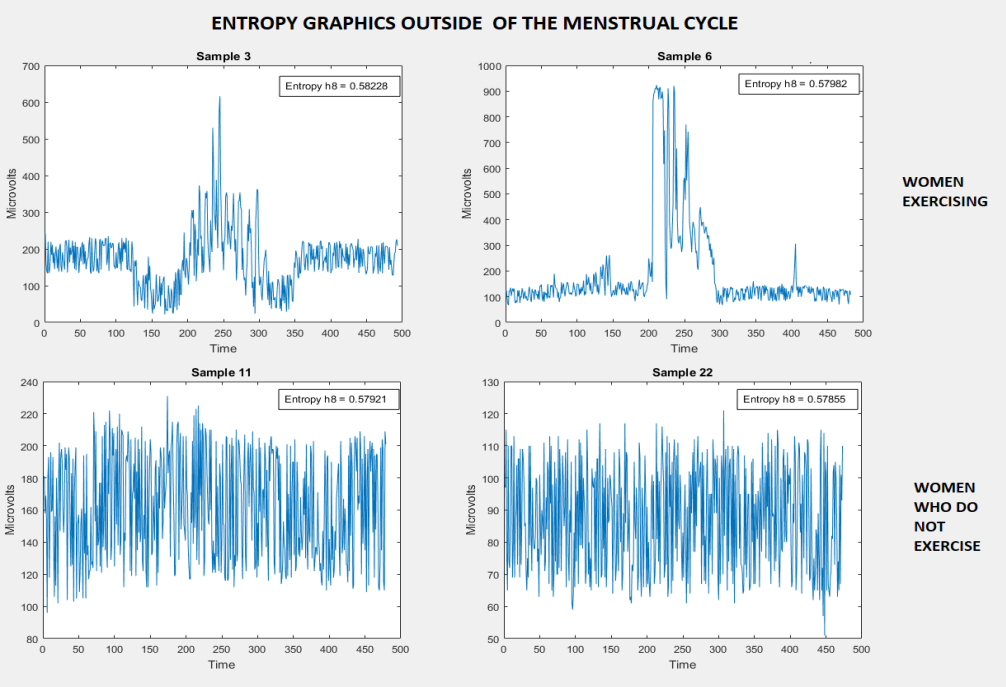


Fig. 6. Some graphics that shows the women without period, and evaluation with entropy being 8 permutation. In the first row are some women graphics exercising, meanwhile in the second row are some women graphics who do not exercise.

6); and to the application of the permuted entropy, in which 8 permutation points were used [13,15].

The entropy used allows the evaluation of nonlinear time series (chaotic series such as the case of the signals obtained in the muscle, figures 5-6), as well as allowing the comparison between signals, thus providing the necessary information to group the elements that belong to the same study.

The Hilbert Huang transform was also applied [20, 22-25], as another comparative method, where for this case study, where the woman is in her period, the forms of the signal are like rose petals, unlike when they are not in their period, the signals are replicated almost in the same way, as shown in figures 7-8. Finally, the Fourier transform [17-19] was applied, with which the enveloping characteristic of a biological signal mentioned by Hodgkin-Huxley in [24-25] is visualized.

IV. Conclusions and Result

It was found that in a series of biological time obtained from the muscular response can be modified by the phase of the

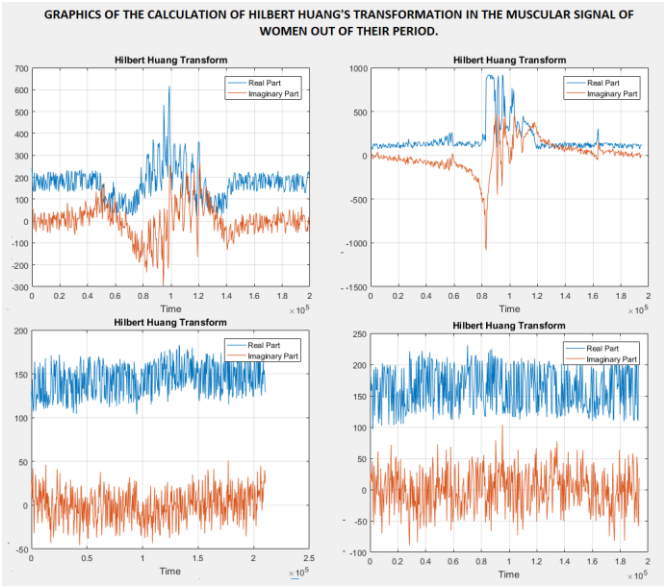


Fig. 7. Graphics of the calculation of Hilbert Huang's transformation in the muscular signal of women out of their period. The red signal is the result of the Hilbert Huang's Transform, and the blue signal is the bio-signal real.

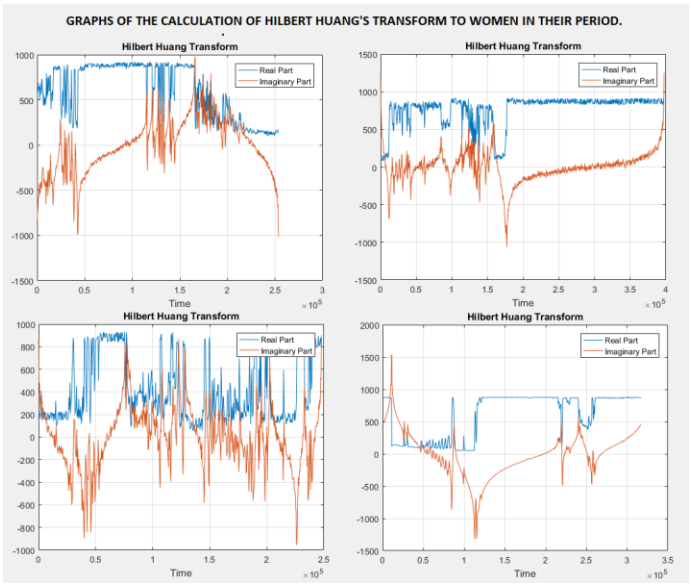


Fig. 8. Graphs of the calculation of Hilbert Huang's transform to women in their period. The red signal is the result of the Hilbert Huang's Transform, and the blue signal is the bio-signal real.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

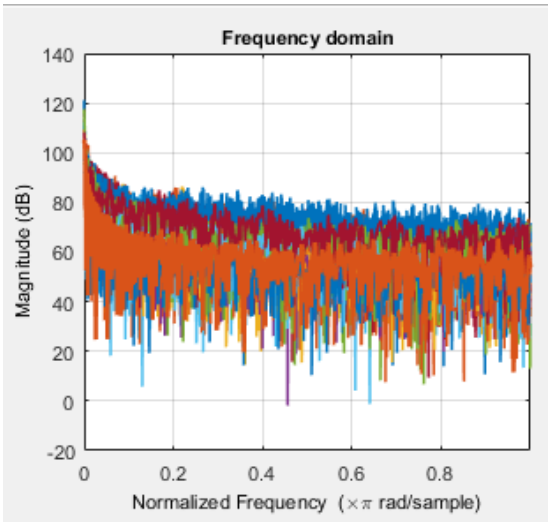


Fig. 9. All individual graphs of the Fourier transform of each bio-signal. Graph of the dominant frequency in the signs of the four weeks of the 7 women.

menstrual cycle, and that this is firstly clearly seen from the data obtained with the electromyography.

Later with the entropy we could know the level of uncertainty of the same signal, and that the singular behavior in the biological signal in the period of a woman, does not distinguish between whether or not exercise is done.

In the results we observed that when the woman is not in her period, between 0.5 - 0.587 entropy, and when she is in the menstrual cycle round between the 0.601 - 0.70 of entropy, always these evaluated with entropy of 8 permutations.

ACKNOWLEDGMENT

Authors would like to acknowledge to CONACYT, the Tecnológico Nacional de México, Instituto Tecnológico de León and our drawer Arlette Rodríguez; their support for this research.

REFERENCES

[1] GARCÍA, José María Artero; HURLÉ, J. Anatomía Y Fisiología Humana. Everest, (1978).
[2] SALAZAR, Osiel Arbeláez; ANGARITA, Jorge Gómez; VARGAS, Jairo Mendoza. Diseño De Un Electromiógrafo Con Procesador Digital De Señales Para Captación De Señales Musculares. Revista Médica De Risaralda, (2007), Vol. 13, No 1.
[3] H. A. Romo, J. C. Realpe And P. E. Jojoa, "Análisis De Señales EMG Superficiales Y Su Aplicación En Control De Prótesis De Mano," Vol. 14, No. 1, (2007).
[4] NEUMANN, Donald A. Cinesiología Del Sistema Musculoesquelético. Paidotribo, (2007).
[5] Jain, M. A Next-Generation Approach To The Characterization Of A Non-Model Plant Transcriptome. Curr. Sci. 101, 1435–1439 (2011).
[6] Ramakrishnan, S. & Karthick, P. A. Muscle Fatigue Analysis Using Surface EMG Signals And Time–Frequency Based Medium-To-Low Band Power Ratio. Electron. Lett. 52, 2–3 (2015).
[7] Al-Mulla, M. R., Sepulveda, F., Colley, M. & Kattan, A. Classification Of Localized Muscle Fatigue With Genetic Programming On Semg During Isometric Contraction. Proc. 31st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. Eng. Futur. Biomed. EMBC 2009 2633–2638 (2009). Doi:10.1109/IEMBS.2009.5335368
[8] Gerardo, J. Et Al. Polynomial Approximation Of Time Series Of Pupil Response To Controlled Light Stimuli. 7, 1–10 (2017).
[9] BRILLINGER, D. R. Time Series: Data Analysis And Theory, Society For Industrial Mathematics (2001).
[10] N. Bhadra, K. L. Kilgore, "High-Frequency Electrical Conduction Block Of Mammalian Peripheral Motor Nerve", Muscle Nerve, Vol. 32, No. 6, Pp. 782-90, Dec (2005).
[11] Liu, H. Steady State Membrane Potential And Sodium Current Changes During High Frequency Electrical Nerve Stimulation. 1006–1009 (2015). Doi:10.1109/ICMTMA.2015.245.
[12] Baxi, K. And Martin, R. Entropia. Architecture, Art, Design, Fashion, History, Photography, Theory And Things. Black Dog Pub. (2001).
[13] Bejan, A. Entropy Generation Minimization: The Method Of Thermodynamic Optimization Of Finite-Size Systems And Finite-Time Processes; CRC Press: Boca Raton, FL, USA, (1995).
[14] B. Christopher, P. Bernd. "Permutation Entropy – A Natural Complexity Measure For Time Series". Procedia Environmental Sciences. Vol. 88. (2002). Pp. 4102-4106.
[15] Qing, J., Bhatti, M. M., Abbas, M. A., Rashidi, M. M. & Ali, M. E. Entropy Generation On MHD Casson Nanofluid Flow Over A Porous Stretching / Shrinking Surface. Doi:10.3390/E18040123
[16] Mosquera-Dussan, O. L. & Botero-Rosas, D. A. Nonlinear Analysis Of The Electroencephalogram In Depth Of Anesthesia Análisis No Lineal De La Señal De Electroencefalograma En Profundidad Anestésica. 45–56 (2015). Doi:10.17533/Udea.Redin.N75a06
[17] E.O. Brigham, The Fast Fourier Transformi, Prentice-Hall, Englewood Cliffs, N. J., (1974).
[18] A. Oppenheim Y R. Schafer, Discrete-Time Signal Processing, Cap. 8, 9 Y 12, Prentice-Hall Inc., (1989).
[19] W.T. Cochran, What is the Fast Fourier Transform?, IEEE, Trans. Audio Electroacoust., Vol. AV-15, (1967), pp. 45-55
[20] 1. Ding, H., Huang, Z., Song, Z. & Yan, Y. Hilbert – Huang transform based signal analysis for the characterization of gas – liquid two-phase flow. 18, 37–46 (2007).
[21] Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of Royal Society London Series A (1998);454:903–95.
[22] Balocchi R, Menicucci D, Varanini M. Empirical mode decomposition to approach the problem of detecting sources from a reduced number of mixtures, EMBS. In: Proceedings of the 25th annual international conference of the IEEE, vol. 3. (2003). p. 2443–6.
[23] Wu ZH, Huang NE. A study of the characteristics of white noise using the empirical mode decomposition method. Proceedings of Royal Society London Series A (2004); 460:1597–611.
[24] Jane Cronin, Mathematical Aspects of Hodgkin-Huxley Neural Theory, Cambridge Studies in Mathematical Biology, Cambridge University Press, 1987; ISBN: 0521334829,9780521334822.
[25] Hodgkin A.L., Huxley A.F. A Cuantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve. J Physil. 117, 500-544 (1952)

Prostate Near-Infrared Fluorescent Image Segmentation Using Hopfield Neural Network Classifier

Rachid Said Sammouda
Computer Science Department
College of Computer and Information Sciences, King Saud University
Riyadh, KSA
dr.rachidsammouda@gmail.com, rsammouda@ksu.edu.sa

***Abstract*— Thousands of people die every year due to prostate cancer. The prognosis of prostate cancer is very slow in most of the cases, but it can cause the death of the patient. The diagnostic pattern and the strategies of health care systems have changed over the last ten years. This change occurred rapidly due to the easy availability and an outburst in patient's data. This data is used as input data to Computer-Aided diagnosis systems. The objective of this research is to improve the diagnosis by developing a prototype arrangement for revealing, detecting and classifying the prostrate tumor. This is achieved by using Near-infrared and Mid-infrared spectrums of prostate pathological images. This optical imaging technique is a potent tool for cancer investigation that relies on stimulating endogenous chromophores or applying contrast agents able to target cancer cells. Here, we present a segmentation method of images obtained using PSMA (Prostate Specific Membrane Antigen) targeting optical imaging probes for NIRF (Near Infrared Fluorescence). This phenomenon is applied for intra-operative visualization of prostate cancer. An Artificial Neural Network classifies the pixels into distinguished clusters. Preliminary tests were conducted. The outcomes of these tests reveal that the planned segmentation technique can enhance the existing clinical practice in identifying prostate area. According to the NIRF image, shape and volume analysis could be conducted using the segmentation result for further investigations.**

Keywords- Hopfield Neural Network Classifier; Near-Infrared Fluorescence optical images; Prostate Cancer; PSMA (Prostate Specific Membrane Antigen); Segmentation;

I. INTRODUCTION

Cancer or malignant tumor occurs due to the abnormal growth of the cells. The prognosis of cancer occurs due to the movement of cancerous cells in the body by using mediums i.e. the blood and lymph. These cancer-cells attack the healthy cells and destroy them. The cancer cell grows by cell division process causes angiogenesis, i.e. formation of new blood vessels. For global public health, cancer has become a major risk factor. Regardless of the progress in wide-ranging therapy, cancer is a straining financial difficulty for patients in all societies. The detection of cancer is very important at its earliest stages. This early detection is very difficult because of the reduced level of specificity and sensitivity regarding current diagnostic approaches of imaging. There are different types of cancer; prostate cancer is one of them [1, 2].

Prostate cancer is the cancer that occurs in the tissues of the prostate gland. The function of prostate gland is the production of seminal fluid. The seminal fluid is required for the nourishment and transportation of sperms. Prostate cancer lives as it is born slow-growing and benign or fast-growing and dangerous 3. The early stage diagnosis of cancer is very important to prevent its prognosis [3]. For this purpose, the research is going on to device new techniques for early diagnosis and detection.

Clinical organizations are working on the prevention and treatment of the cancer. Furthermore, different strategies are planned to improve the diagnostic methods. The new aim is to develop the non-invasive methods such as imaging method. The biomedical imaging devices are used very frequently today, while further developments are underway to produce more advanced apparatus. These devices work at the cellular, molecular or tissue levels and make the diagnosis more accurate and favorable. The studies at the molecular and cellular levels help to know the mechanism of the prognosis of cancer. In addition to this, the patients prefer the non-invasive method of diagnosis. Keeping all these points in mind, imaging methods are becoming common and advancements in imaging modalities are progressing. The appropriate usage of near infrared segmentation is also a new approach for diagnosing the condition of prostate cancer. The studies are carried out on mice to evaluate the affectivity of the proposed method [4, 5].

The extensively used method for the detection of the pathological changes occurring due to cancer is imaging modalities. The examples of these widely used imaging methods are CT, PET, ultrasound, and MRI. Such methods show results in cases of benign lesions. However, in malignancies the imaging technique fails to get a clear contrast between the benign and the malignant. Moreover, the adjacent normal tissues add further confusion. To improve the detection and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

examination of cancer at any phase, it is very essential to develop a high contrast narrative imaging method to augment the diagnosis and therapeutics [1].

Prostate cancer and breast cancer are two frequent types of cancer and their diagnosis has become a challenge for the scientists. The difficulty in localization of cancer cells is the main barrier. Furthermore, it is very difficult to differentiate between the normal and tumor cells. New strategies are designed to localize the cancer cells in imaging. These strategies comprises of the techniques of labeling methods. The ligands which are tumor-specific and having sympathetic pharmacokinetics are developed for labeling purpose [1].

The method of capillary permeability and in vivo tumor growth selectively increase the expression of tumor markers, and tumor delivery changes. Favorable pharmacokinetic, and small tumor markers developed using pre-targeting strategies are important in improving the diagnostic approach [2]. In diagnosing cancer, optical imaging by Near-infrared Fluorescence (NIRF) is a dominant trend. It relies on activating endogenous chromophores or applying contrast agents that can target cells. Several new NIRF agents have been developed including heptamethine carbocyanine dyes. Some of these agents have become commercially available in recent years, such as Cy5.5 [4] and IR Dye 800-CW [5]. These have been coupled with peptides or antibodies and successfully used for the targeted visualization of neoplastic tumors in animal models [6]. Xinning et al. [7] and others [8-12] have developed optical molecular imaging approaches to differentiate between tumors and surrounding normal tissues during surgery, for reviews see [13-15]. A first-in-human-study has been conducted for ovarian cancer [16], indicating progress in this field. The clinical and medical functions of these new NIRF agents offer great promise for future.

The over expressed antigen in most of prostate cancer is PSMA [17-22]. This is the reason that it is a useful biomarker for discrimination of prostate cancer tissue from surrounding normal tissues. Prostate tumor expressing the PSMA receptors were implanted into the flank of mice as previously described [23]. Control tumors that did not express the PSMA receptor were implanted on the opposite flank. When tumors reached the appropriate size the mice were administered a ligand for PSMA labeled with a fluorophore for detection by fluorescence imaging.

II. Generation of Data

A. Mouse Tumor Xenograft Models

Animals were observed every other day until tumors reached about 10 mm in diameter. Orthotropic implantation of prostate cancer was carried as previously described. Briefly, six to eight weeks old male nude mice lacking thymus gland were anesthetized. The composition of anaesthetic solution was 5 mg/mL ketamine/ 3 mg/mL xylazine solution in 0.9% saline and the volume given was 200 uL. The route of administration was intra-peritoneal (in the peritoneal cavity). The lower abdomen was open to expose the dorsal-lateral prostate, to which 10 to 20 uL cell suspension in PBS (5x107 cells/ml) was injected. The incision in the abdominal wall was closed. After four weeks, animals were ready for experimentation.

B. In-vivo NIR Imaging Studies

With the assistance of Maestro in- vivo Imaging system (Perkin-Elmer, Waltham, MA), imaging was performed. 1 nmol of NIR probe in PBS through tale vein injection was given to each mouse. Imaging was carried out by using the appropriate filter set (deep red filter set for PSMA-1-IR800 and yellow filter set for PSMA-1-Cy5.5). Different points were selected for imaging. The temperature of 37oC was tuned for the imaging bed during imaging. A nose cone was adjusted with imaging bed for inhalation of isofluorane. Cervical dislocation was used to sacrifice mice after imaging mice over 5 days post injection. To perform ex vivo imaging, harvesting of tissues, for example kidneys, heart bladder and liver was done.

Fluorescent molecular tomographic (FMT) images were obtained using the FMT2500 device (Perkin-Elmer, Waltham, MA) and three-dimensional reconstructions of fluorescent signals were acquired using the accompanying software, TrueQuant. Quantification of fluorescent signals was obtained by calibration of PSMA-1-IR800 and PSMA-1-Cy5.5 using the 780 nm and 680 nm channel respectively. To block the binding of PSMA-1-NIR in mice, mice were co-injected with 1 nmol of PSMA-1-NIR probes and 100 nmol of ZJ-MCC-Ahx-YYYG, an analogue of PSMA-1 with similar binding affinity but with no optical probe attached.

Maestro Imaging System and FMT were the two imaging methods that were used to image mice for up to 24 hours. For orthotopic mouse models, mice were imaged at 4 hours or 24 hours by using Maestro Imaging System. 1 nmol of PSMA-IR800 injection at 4 hours or 1 nmol of PSMA-1-Cy5.5 at 24 hours were injected in post tail vein. After the completion of the optical imaging, the mouse was euthanized, the abdomen was opened to expose the tumor, and the mouse was again imaged. Finally, tumor was harvested for ex vivo imaging.

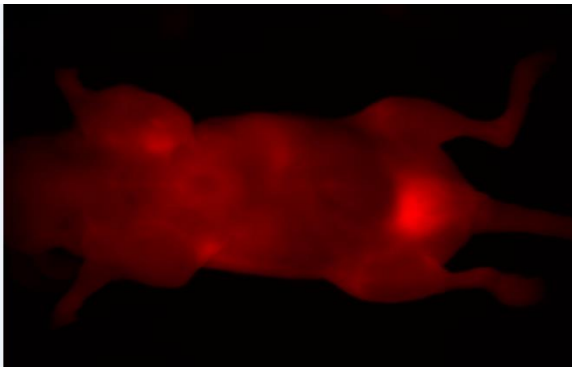


Figure 1. Shows a NIRF sample image of mice model with prostate cancer.

Figure1 shows a NIRF image of prostate tumors obtained by using targeted imaging probe of Prostate specific membrane antigen in a mouse model. Several similar NIRF images were collected in our previous study which has been conducted to develop PSMA-targeted near infrared (NIR) optical imaging probes. These were used for visualization of prostate cancer intra-operatively. A high affinity PSMA ligand (PSMA-1) was

National Plan for Science and Technology at King Saud University (KSU)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

synthesized with low molecular weight, and further labeled with commercially available NIR dyes: IRDy800 and Cy5.5 [5]. It demonstrated the utility of such probes to selectively bind to prostate tumor in vivo targeting both heterotopic and orthotropic prostate tumors. A challenge for these types of studies is to correctly interpret the imaging data to accurately reflect the margin of the cancer. In cancer research, it is very difficult to obtain reproducible, accurate, precise and intent assessment. The dilemmas occur due to the variability of personnel, biological dissimilarity, and natural unpredictability. NIRF imaging technique identifies the cancer by providing fluorescent information from every pixel in the image. For prostate cancer, the NIRF imaging CAD system (Computer Aided Diagnosis) could be classified and analyzed to build a set of sharp diagnostic rules. We present a segmentation method of the NIRF images as the first and bottleneck entity of the CAD system in the next section

III. Segmentation Method

All of the above described phenomena were studied to determine the importance and use of infrared radiation for obtaining a better health approach [24]. To screen medical imaging, segmentation of image is imperative. A progressive method used for screening purposes in the last few decades is fuzzy segmentation method. The broadly used fuzzy method is based on c-means algorithm. The accomplishment of introducing fuzziness for each image pixel is successful as it is fit for images. The fuzziness promotes the bunching and clustering of the image pixels. This method assists in preserving more information in cluster form. The original image obtained has hard and crisp segmentation process which does not give precise information. That is why the clustering method is preferred [25].

Similar to that, we have used the Unsupervised Hopfield Neural Network Classifier (UHNNC) in segmentation of different types of medical and natural color images [26, 27]. The segmentation results have been appreciated with respect to the multi-dimensionality of the data type used for segmentation. This means the UHNNC gives better segmentation results as far as getting more information about the pixel of the scene under segmentation.

A grid of NM neurons is present in UHNNC architecture. The rows and columns are well defined. In NM grid, the alphabet N is used to show the size of the image; whereas the numbers of the cluster formed are represented by M. Columns are used to characterize a class while pixels are represented in row form. The network is deliberate to sort the area of the features themselves.

By using a distance scale, compactness of each category is calculated. The problem of Segmentation is considered as a partition of N pixels of P characteristics among M clusters or so that the cost of energy (errors) function can be minimized by the tasks of pixels:

$$E = \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{M} R_{kl}^{n} V_{kl}^{2} \quad (1)$$

The similarity distance is measured and represented by R_{kl}. It shows the distance between k[th] pixel and the centroid of class *l*. It is given as:

$$R_{kl} = \|X_k - \bar{X}_l\| \quad (2)$$

In the above equation X_k represents the P-features vector, for color images, P=3 in the RGB color space, as k[th] pixel's intensities, while $\bar{X}_l$ is the class l's centroid, and is shown as:

$$\bar{X}_l = \frac{\sum_{k=1}^{N} X_k V_{kl}}{n_l} \quad (3)$$

To allocate a label m to the pixel, the input-output function for the k^{th} row, winner-takes-all learning is used by HNN. It is given by:

$$\begin{cases} V_{kl}(t+1) = 1, \ if \ \ U_{kl} = Max\{U_{kl}(t), \forall\, l\} \\ V_{kl}(t+1) = 0; \qquad\qquad otherwise \end{cases} \quad (4)$$

UHNNC is used for the minimization purpose and by working out a group of equations of motion, the resultant obtained is:

$$\frac{\partial U_i}{\partial t} = -\mu(t)\frac{\partial E}{\partial V_i} \quad (5)$$

U_i , represents the input of i^{th} neuron, while the output is represented by symbol V_i. For increasing the convergence speed of the HNN, $\mu(t)$ is used as a scalar positive function of time that we have defined and verified in our study its efficacy in assuring and forcing the network to converge after a pre-specified time T_s as follows:

$$\mu(t) = t * (T_s - t) \quad (6)$$

A group of neural dynamics is obtained by relating the equation (5) to equation (1) and is given by:

$$\frac{dU_{kl}}{dt} = -\mu(t)(R_{kl}^{n} V_{kl}) \quad (7)$$

The UHNNC segmentation algorithm can be easily summarized as:

Phase1. The neurons' input is initialized to randomly assigned values.

Phase2. The new output value for every neuron can be obtained by applying the input-output relation given in (4).

Phase3. With respect to the equation (3), centroid can be computed for each class.

Phase4. The input of each neuron is required to be updated by solving the set of differential equation in (5) and given as:

$$U_{kl}(t+1) = U_{kl}(t) + \frac{dU_{kl}}{dt} \quad (8)$$

Phase5. Loop to Phase2 with T_s times.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. SEGMENTATION RESULTS

The main objective regarding the preliminary research is to reveal the importance of the anticipated approach of segmentation for the diagnosis of prostate cancer by using NIRF imaging technique. However, although optical imaging by NIRF in cancer researches is an influential research instrument as mentioned above, it remains a one dimensional information set about the scene's environment and its images' segmentation using UHNNC which is of limited contrast.

To overcome that contrast limitation, we have produced an artificial multidimensionality of the NIRF image using dependent chromatic redundancy in the RGB color space. Figure 2 (a) shows the NIRF sample image of mice model with prostate cancer of Figure1.

The Green and Blue channels obtained by redundancy from Figure 2 (a) are shown in (b) and (c) parts of Figure 2. The (d) part of Figure 2 regarding the RGB color space shows the full color display of the three components. The above described UHNNC is applied to several NIRF images. The results show that segmentation of most of the images can be obtained successfully by using our algorithm. The segmentation has clearly distinguishable areas as background, or other uniform clusters with respect to their features in the input images.

Figure 3 shows the segmentation result using the UHNNC of the NIRF image of a mice model Figure2 (a) and its two redundant green and blue color filters, Figure 2 (b) and (c), with respect to number of clusters, 3, 4, 5, and 6, respectively to (a), (b), (c), and (d).

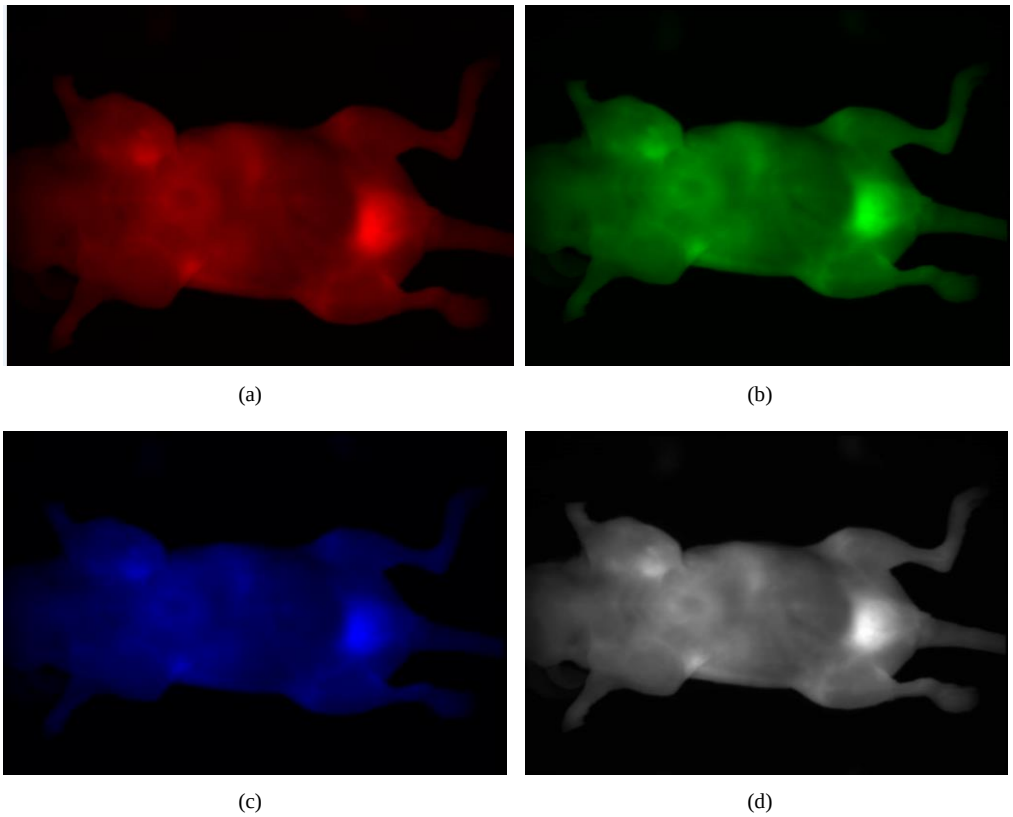


Figure 2. (a) is the raw NIRF image of the case under study, the same image in Figure 1, (b) and (c) are green and blue component obtained by redundancy from (a), and (d) is the full color display of (a), (b), and (c) with respect to the RGB color space.

The curves of the segmentation problem energy function are shown in Figure 4. The segmentation problem during its optimization using the described UHNNC with respect to the number of cluster L is decided by the user based on anatomical, medical information.

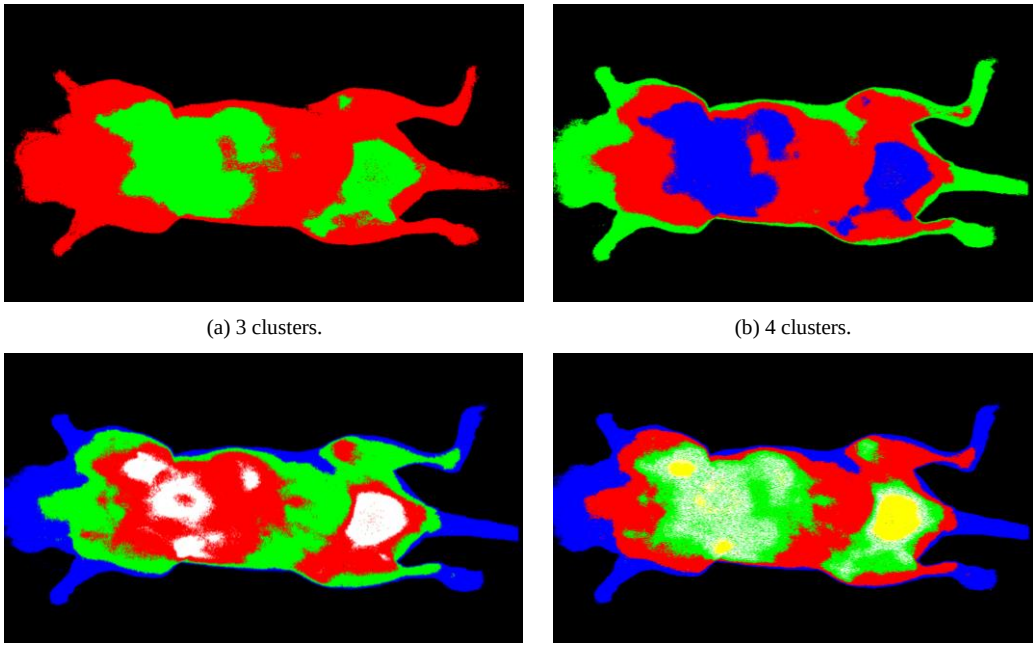
We realize that the prostate region starts to appear as independent region with its outer borders when clusters number is equal to five.

Figure 6 (d) obtained with six clusters, shows the prostate region with an outer and inner regions.

Figure 5 shows more specific regions within the prostate area, for which we have conducted segmentation with more clusters, eight, nine and ten. The convergence optima of the UHNNC with respect to the number of clusters used during the segmentation process is shown in Figure (6).

As can be seen, the UHNNC has reached better local optima when used with ten clusters, as there are more intensity variations among pixels, however, the prostate region, cluster number 4 in Figure 7, remains among the cluster of the highest mean value.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(a) 3 clusters. (b) 4 clusters.

(c) 5 clusters. (d) 6 clusters.

Figure 3. shows the segmentation result using the UHNNC of the NIRF image of a mice model Figure2 (a) and its two redundant green and blue color filters, Figure 2 (b) and (c), with respect to number of clusters, 3, 4, 5, and 6, respectively to (a), (b), (c), and (d).

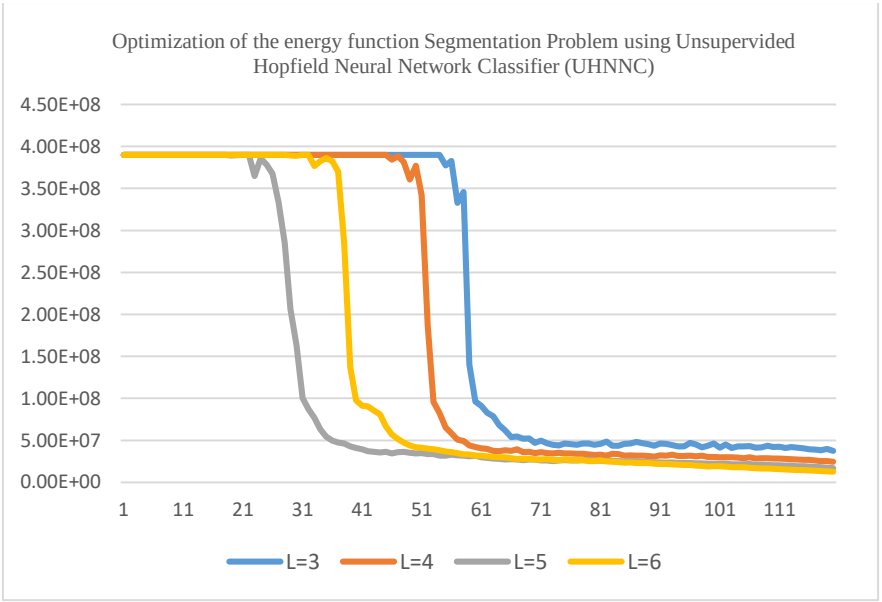
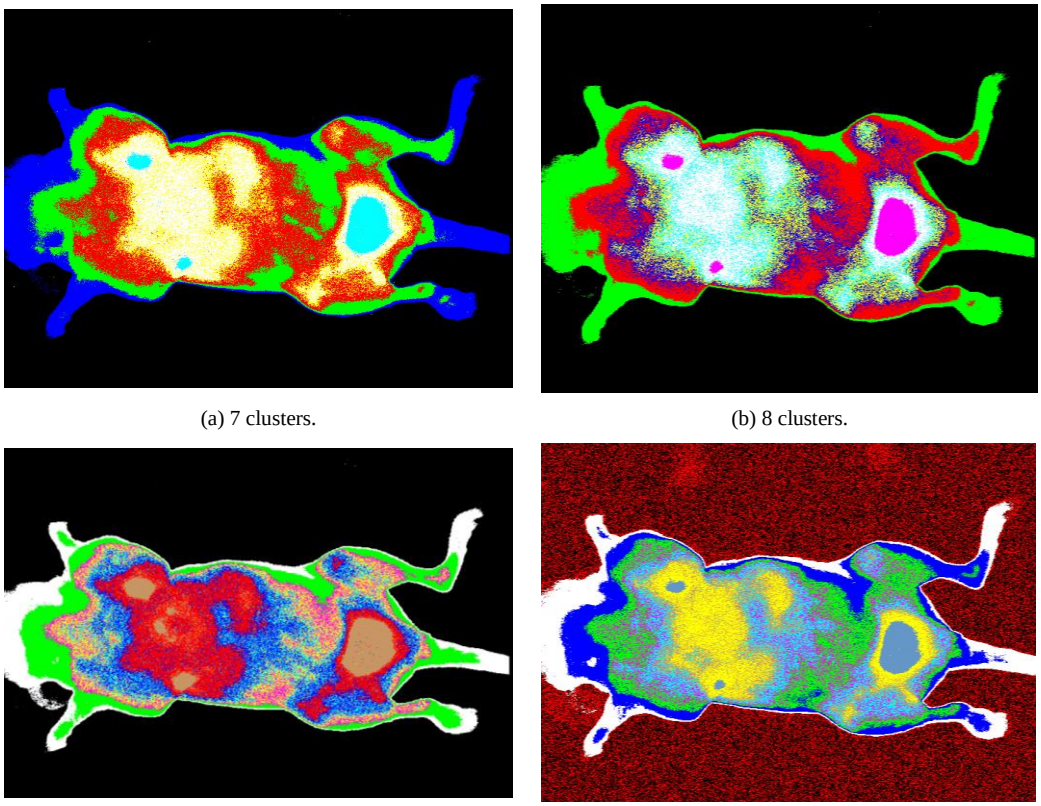


Figure 4. shows the curves of the segmentation problem energy function during its optimization using the here described UHNNC with respect to the number of cluster, L, decided by the user.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(a) 7 clusters. (b) 8 clusters.

(c) 9 clusters. (d) 10 clusters.

Figure 5. shows the segmentation result using the UHNNC of the NIRF image of a mice model Fig. 2 (a) and its two redundant green and blue color filters, Fig. 2 (b) and (c), with respect to number of clusters, 7, 8., 9 and 10, respectively to (a), (b), (c), and (d).

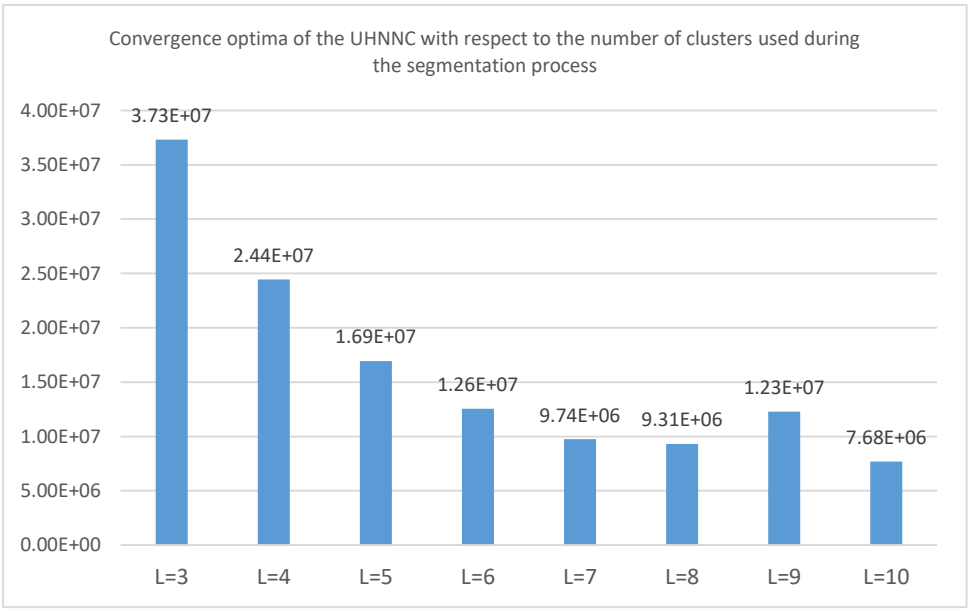


Figure 6. shows the convergence value of the energy function of the UHNNC during the segmentation process, of the NIRF image shown in Fig. 1, with respect to the number of clusters decided by the user.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

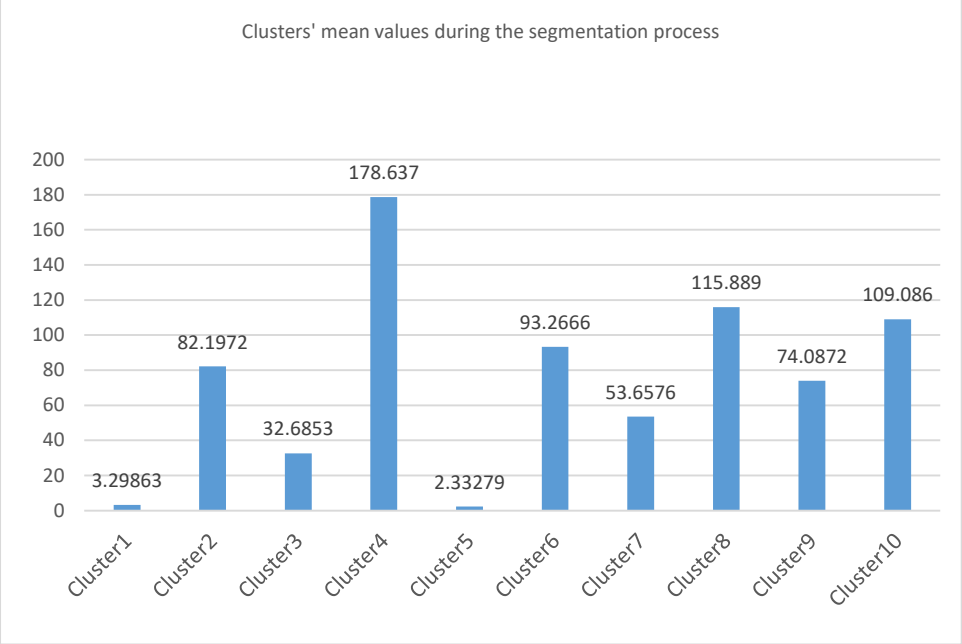
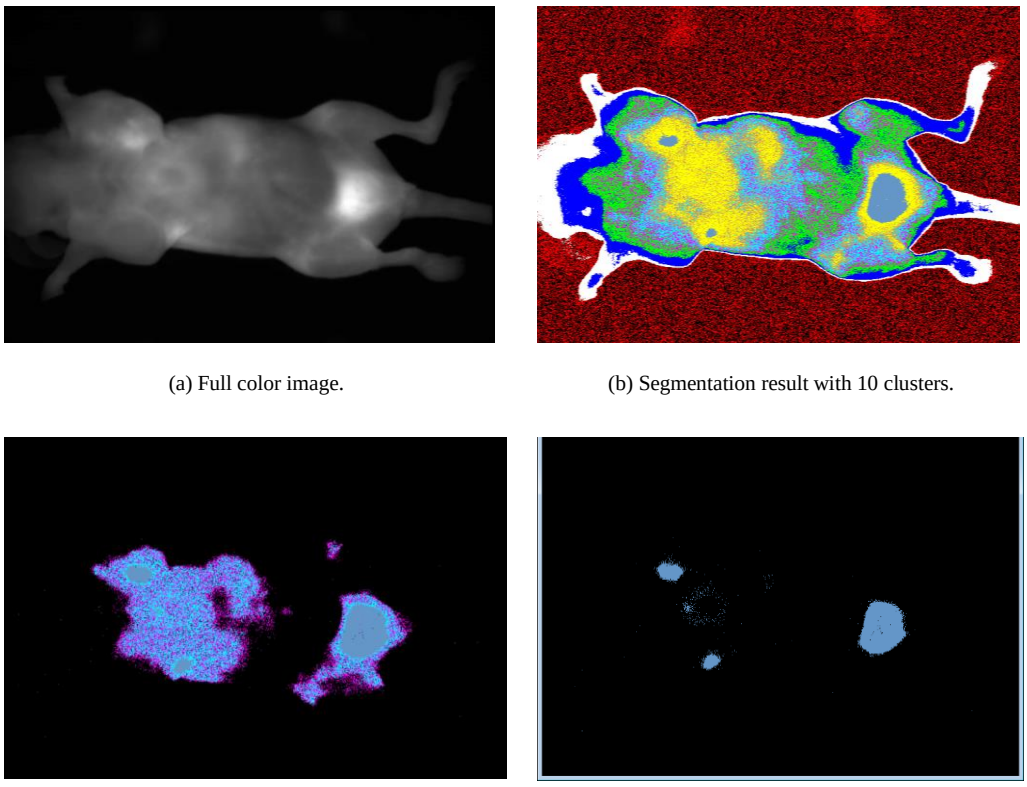


Figure 7. shows the mean value of each cluster in the segmentation result of the NIRF image shown in Fig. 1, with 10 as number of clusters, decided by the user.



(a) Full color image.

(b) Segmentation result with 10 clusters.

(c) Clusters with the first three maximums of mean values.

(d) The cluster having the maximum mean value.

Figure 8. shows the full color image of the case under study in (a), its corresponding segmentation result with 10 clusters in (b), the clusters of the first three maximum mean values in (c) and the cluster with the maximum mean value, including the prostate region, in (d).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 8 shows the details of the full color image of the case under study in (a). While in (b) the result with 10 clusters of corresponding segmentation is shown. The clusters of the first three maximum mean values in (c) and with the maximum mean value, including the prostate region, is shown in (d).

We can realize here that even with more clusters, the prostate region remains as an entity region with lower intensity variation, and did not split into two clusters as the background. The latter will be used as a mask to take out the section of attention from the unrefined image for further diagnosis and design of the CAD system for prostate cancer diagnosis. All these figures make the procedure of segmentation of near-infrared fluorescent easy to understand.

V. Discussion

The imaging obtained by using near infrared is very helpful for the diagnosis as well as for the surgical approach. It provides accurate images of the cancer cells that differ from the images of normal cells. The differentiation makes the diagnosis easy and painless. The differentiation marked by using this technique aids the dissection and categorization of the tumor related cells without developing any harmful effect [16].

This near-infrared imaging technique not only acts as a diagnostic tool, but also traces the response of the cells to the chemotherapeutic agents. For the therapy of prostate cancer, it is very important to formulate a drug which has high therapeutic efficacy and fewer side effects. This is done by taking the measurements of the images. The measurements are taken to find out the reduction in the size of tumor. This, however, is a lengthy process. Despite long time delay, this technique is considered as an important indicator for the trial of new drugs. The efficacy of a new drug molecule can be determined by using this approach. Thus this tool is helpful in selecting the most significant and effective treatment [27].

One of the recent modality of imaging is hyper-spectral imaging. It is a spectroscopic method, and the data obtained from this imaging method is utilized for non-invasive approach in cancer detection. The differentiation of tumor cells from the healthy cells is necessary. This is done by quantitative analysis. For prostate cancer, the data is obtained by use of an advanced image data. The analysis of hyper-spectral image is done to obtain the data which is utilized for the detection of cancer. For the purpose of differentiation of normal and cancerous cell, the spectrum was taken out for both kinds of cells. The studies were conducted in order to detect prostate cancer on the mice having tumor. Moreover, pathological slides were also used for detection. By using this technique, the images of normal and tumor cells were taken and the reflectance properties of both cells were extracted. These images showed that the reflectance properties of both cells are different. The sensitivity and specificity of this method are fine. By using the data obtained by spectral images, is very helpful to differentiate between normal and tumor cell, so the safe dissection of malignant areas is possible [28]. The determination of the in vivo cell death is possible by the use of near infrared fluorescent method. For this purpose, fluorescent probes are used. For example, active Cy-annexin is used in non-radioactive techniques. NIRF probe having active cy-annexin is used to determine the anti-proliferative properties of the molecules which are used as chemotherapeutic regimens. By analyzing the properties of the regimens, it is very easy for the clinicians to choose the chemotherapeutic agents for the prostate cancer [28].

Another study indicates the importance of the near-infrared spectroscopy for the diagnosis and detection purpose. The near infrared image segmentation has achieved the significance because of its non-invasive property. Due to this property, the technology is widely used. The diagnostic markers are used to detect the chromophores difference [29].

The endogenous chromophores of normal cells are different from the cancer cells, and this quality is used to identify the cancer cells. This detection is based on near-infrared radiations and the biomarkers (for example, lipids bands, deoxy-haemoglobin, oxy-haemoglobin and water bands etc). In addition to NIR, different agents are used to increase the contrast of the image [30]. All these studies indicate that near infrared rays fluorescent imaging is a very important method for the detection of the cancer cells. The segmentation of the image obtained by this technique is an advanced approach for this procedure. This new innovation is very promising and it can be used as a potential aid in the war against prostate cancer. In the future work, we will apply the method proposed in [31] for NIRF images de-noising before segmentation them by the previous proposed methods in [32-34] and compare the results with the proposed method in this paper.

VI. Conclusion

In this paper, we presented the use of NIRF images for prostate cancer diagnosis. NIRF technology has been considered widely for biomedical research and clinical application since it has been demonstrated that near-infrared is an appropriate optical opening for profound tissue imaging.

As shown in the sample of NIRF images, used in this study, these dyes show the ability of providing fine information and behaviors about different mice's tissues. The finest information is utilized to develop the new strategies of diagnosis and detection. The segmentation of these NIRF images using our modified UHNNC confirms the fact that prostate cancerous tissue takes more fluorescent material than normal tissue.

The analysis process conducted among the different clusters, of the segmentation results; prove the low intensity variation among pixels of the cancerous tissue. This makes the prostate cancerous region presented by smooth region and sharp edges. In our future work, we will use these features in order to extract automatically the region of interest (ROI) as prostate tissue, and focus on the internal behavior of its cells for better guidance in prostate cancer therapy and early diagnosis.

Acknowledgment

The author is appreciative to National Plan for Science and Technology for their assistance in this research at King Saud University (KSU), Riyadh, Kingdom of Saudi Arabia (KSA), under project number: 10-Bio-1905, and the hosting of the assignment provided by the College of Computer and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Information Sciences at KSU. Also, the author is appreciative to James P. Basilion researcher at the Department of Radiology and NFCR Center for Molecular Imaging, Case Western Reserve University, and Cleveland, OH, USA for his support as consultant in this project.

REFERENCES

[1] X. Yi, F. Wang, W. Qin, X. Yang and J. Yuan, "Near-infrared fluorescent probes in cancer imaging and therapy: an emerging field", International Journal of Nanomedicine, p. 1347, 2014.

[2] R. Weissleder, C. Tung, U. Mahmood and A. Bogdanov, "In vivo imaging of tumors with protease-activated near-infrared fluorescent probes", *Nature Biotechnology*, vol. 17, no. 4, pp. 375-378, 1999.

[3] D. Pendick, "Prostate cancer lives as it is born: slow-growing and benign or fast-growing and dangerous - Harvard Health Blog", *Harvard Health Blog*, 2017. [Online]. Available: http://www.health.harvard.edu/blog/prostate-cancer-lives-as-it-is-born-slow-growing-and-benign-or-fast-growing-and-dangerous-201308146604. [Accessed: 21- Dec- 2017].

[4] S. Banerjee, M. Pullambhatla, Y. Byun, S. Nimmagadda, G. Green, J. Fox, A. Horti, R. Mease and M. Pomper, "68Ga-Labeled Inhibitors of Prostate-Specific Membrane Antigen (PSMA) for Imaging Prostate Cancer", *Journal of Medicinal Chemistry*, vol. 53, no. 14, pp. 5333-5341, 2010.

[5] Y. Chen, S. Dhara, S. Banerjee, Y. Byun, M. Pullambhatla, R. Mease and M. Pomper, "A low molecular weight PSMA-based fluorescent imaging agent for cancer", *Biochemical and Biophysical Research Communications*, vol. 390, no. 3, pp. 624-629, 2009.

[6] J. Wu, D. Pan and L. Chung, "Near-infrared fluorescence and nuclear imaging and targeting of prostate cancer", *Amepc.org*, 2017. [Online]. Available: http://www.amepc.org/tau/article/view/2764/3637. [Accessed: 21- Dec- 2017].

[7] J. Cutter, N. Cohen, J. Wang, A. Sloan, A. Cohen, A. Panneerselvam, M. Schluchter, G. Blum, M. Bogyo and J. Basilion, "Topical Application of Activity-based Probes for Visualization of Brain Tumor Tissue", *PLoS ONE*, vol. 7, no. 3, p. e33060, 2012.

[8] G. Blum, S. Mullins, K. Keren, M. Fonovič, C. Jedeszko, M. Rice, B. Sloane and M. Bogyo, "Dynamic imaging of protease activity with fluorescently quenched activity-based probes", *Nature Chemical Biology*, vol. 1, no. 4, pp. 203-209, 2005.

[9] G. Blum, G. von Degenfeld, M. Merchant, H. Blau and M. Bogyo, "Noninvasive optical imaging of cysteine protease activity using fluorescently quenched activity-based probes", *Nature Chemical Biology*, vol. 3, no. 10, pp. 668-677, 2007.

[10] Q. Nguyen, E. Olson, T. Aguilera, T. Jiang, M. Scadeng, L. Ellies and R. Tsien, "Surgery with molecular fluorescence imaging using activatable cell-penetrating peptides decreases residual cancer and improves survival", *Proceedings of the National Academy of Sciences*, vol. 107, no. 9, pp. 4317-4322, 2010.

[11] Y. Urano, M. Sakabe, N. Kosaka, M. Ogawa, M. Mitsunaga, D. Asanuma, M. Kamiya, M. Young, T. Nagano, P. Choyke and H. Kobayashi, "Rapid Cancer Detection by Topically Spraying a -Glutamyltranspeptidase-Activated Fluorescent Probe", *Science Translational Medicine*, vol. 3, no. 110, pp. 110ra119-110ra119, 2011.

[12] M. Veiseh, P. Gabikian, S. Bahrami, O. Veiseh, M. Zhang, R. Hackman, A. Ravanpay, M. Stroud, Y. Kusuma, S. Hansen, D. Kwok, N. Munoz, R. Sze, W. Grady, N. Greenberg, R. Ellenbogen and J. Olson, "Tumor Paint: A Chlorotoxin:Cy5.5 Bioconjugate for Intraoperative Visualization of Cancer Foci", *Cancer Research*, vol. 67, no. 14, pp. 6882-6888, 2007.

[13] S. Keereweer, J. Kerrebijn, P. van Driel, B. Xie, E. Kaijzel, T. Snoeks, I. Que, M. Hutteman, J. van der Vorst, J. Mieog, A. Vahrmeijer, C. van de Velde, R. Baatenburg de Jong and C. Löwik, "Optical Image-guided Surgery—Where Do We Stand?", 2017. [Online]. Available: http://link.springer.com/article/10.1007/s11307-010-0373-2#page-1. [Accessed: 21- Dec- 2017].

[14] U. Mahmood and R. Weissleder, "Near-Infrared Optical Imaging of Proteases in Cancer", *Molecular Cancer Therapeutics*, 2017. [Online]. Available: http://mct.aacrjournals.org/content/2/5/489.short. [Accessed: 21- Dec- 2017].

[15] "Molecular Imaging | Radiology", *Pubs.rsna.org*, 2017. [Online]. Available: http://pubs.rsna.org/doi/abs/10.1148/radiology.219.2.r01ma19316. [Accessed: 21- Dec- 2017].

[16] G. van Dam, G. Themelis, L. Crane, N. Harlaar, R. Pleijhuis, W. Kelder, A. Sarantopoulos, J. de Jong, H. Arts, A. van der Zee, J. Bart, P. Low and V. Ntziachristos, "Intraoperative tumor-specific fluorescence imaging in ovarian cancer by folate receptor-α targeting: first in-human results", 2017.

[17] R. Israeli, W. Miller, S. Su, C. Powell, W. Fair, D. Samadi, R. Huryk, A. DeBlasio, E. Edwards, G. Wise and W. Heston, "Sensitive Nested Reverse Transcription Polymerase Chain Reaction Detection of Circulating Prostatic Tumor Cells: Comparison of Prostate-specific Membrane Antigen and Prostate-specific Antigen-based Assays", *Cancer Research*, 2017. [Online]. Available: http://cancerres.aacrjournals.org/content/54/24/6306.short. [Accessed: 21- Dec- 2017].

[18] Tasch, M. Gong, M. Sadelain and W. Heston, "A Unique Folate Hydrolase, Prostate-Specific Membrane Antigen (PSMA): A Target For Immunotherapy?", 2017. [Online]. Available: http://www.dl.begellhouse.com/journals/2ff21abf44b19838,46f5c0d0074d488c,6a5b9a2d2ec1da76.html. [Accessed: 21- Dec- 2017].

[19] S. Mannweiler, P. Amersdorfer, S. Trajanoski, J. Terrett, D. King and G. Mehes, "Heterogeneity of Prostate-Specific Membrane Antigen (PSMA) Expression in Prostate Carcinoma with Distant Metastasis", *Pathology & Oncology Research*, vol. 15, no. 2, pp. 167-172, 2008.

[20] J. Trover, M. Beckett and G. Wright, "Detection and characterization of the prostate-specific membrane antigen (PSMA) in tissue extracts and body fluids", *International Journal of Cancer*, vol. 62, no. 5, pp. 552-558, 1995.

[21] J. Ross, C. Sheehan, H. Fisher, R. Kaufman, P. Kaur, K. Gray, I. Webb, G. Gray, R. Mosher and B. Kallakury, "Correlation of Primary Tumor Prostate-Specific Membrane Antigen Expression with Disease Recurrence in Prostate Cancer", *Clinical Cancer Research*, 2017. [Online]. Available: http://clincancerres.aacrjournals.org/content/9/17/6357.short. [Accessed: 21- Dec- 2017].

[22] X. Wang, L. Yin, P. Rao, R. Stein, K. Harsch, Z. Lee and W. Heston, "Targeted treatment of prostate cancer", 2017. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/jcb.21491/abstract;jsessionid=0775C19EECCDB611802293FBCA1F86C5.f01t03?deniedAccessCustomisedMessage=&userIsAuthenticated=false. [Accessed: 21- Dec- 2017].

[23] X. Wang, S. Huang, W. Heston, H. Guo, B. Wang and J. Basilion, "Development of Targeted Near-Infrared Imaging Agents for Prostate Cancer", 2017. [Online]. Available: http://mct.aacrjournals.org/content/13/11/2595.short. [Accessed: 21- Dec- 2017].

[24] X. Liora, R. Reddy, B. Matesic, R. Bhargava, "Towards better human capability in diagnosing prostate cancer using infrared spectroscopic imaging", In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*. ACM July7-11, 2007; pp. 2098-105. Retrieved from, http://dl.acm.org/citation.cfm?id=1277366

[25] X. Yang, Q. Song, Y. Wang, A. Cao and Y. Wu, "A Modified Deterministic Annealing Algorithm for Robust Image Segmentation", *Journal of Mathematical Imaging and Vision*, vol. 30, no. 3, pp. 308-324, 2008. 1

[26] R. Sammouda, J. Hassan and M. Sammouda, "CT images analysis for early detection of lung cancer", *Int. Innov. Comput. Inform. And Control* (IJICIC), vol. 4, no. 11, pp. 2847-2860, 2008.

[27] R. Sammouda, N. Adgaba, A. Touir and A. Al-Ghamdi, "Agriculture satellite image segmentation using a modified artificial Hopfield neural network", *Computers in Human Behavior*, vol. 30, pp. 436-441, 2014.

[28] A. Petrovsky, E. Schellenberger, L. Josephson, R. Weissleder and A. Bogdanov, "Near-Infrared Fluorescent Imaging of Tumor Apoptosis", *Cancer Research*, 2017. [Online]. Available:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

http://cancerres.aacrjournals.org/content/63/8/1936.short. [Accessed: 21-Dec- 2017].

[29] H. Akbari, L. Halig, D. Schuster, A. Osunkoya, V. Master, P. Nieh, G. Chen and B. Fei, "Hyperspectral imaging and quantitative analysis for prostate cancer detection", *Journal of Biomedical Optics*, vol. 17, no. 7, p. 0760051, 2012.

[30] V. Kondepati, H. Heise and J. Backhaus, "Recent applications of near-infrared spectroscopy in cancer diagnosis and therapy", *Analytical and Bioanalytical Chemistry*, vol. 390, no. 1, pp. 125-139, 2007J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[31] R. Sammouda, A. Al-Salman, A. Gumaei and N. Tagoug, "An Efficient Image Denoising Method for Wireless Multimedia Sensor Networks Based on DT-CWT", *International Journal of Distributed Sensor Networks*, vol. 11, no. 11, p. 632568, 2015.

[32] A. AlSalman, A. El-Zaart, S. Al-Salman, and A. Gumaei, "A novel approach for Braille images segmentation", in *Proceedings of the International Conference on Multimedia Computing and Systems (ICMCS '12)*, pp. 190–195, IEEE, Tangier, Morocco, May 2012.

[33] A. Gumaei, A. El-Zaart, M. Hussien, and M. Berbar, "Breast segmentation using k-means algorithm with a mixture of gamma distributions", In *Broadband Networks and Fast Internet (RELABIRA), 2012 Symposium on*, pp. 97-102. IEEE, 2012.

[34] A. Gumaei, A. El-Zaart, and H. Mathkour, "An efficient iris segmentation approach", in International Conference on Graphic and Image Processing (ICGIP '11), vol. 8285 of Proceedings of SPIE, Cairo, Egypt, September 2011.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

New Image Processing Techniques Using Elitism Immigrants Multiple Objective of Genetic Algorithms for Disease Detection

Khalil Ibrahm Mohammed Abuzanounneh
IT Department, College of Computer, Qassim University Saudi Arabia
ka.abuzanouneh@qu.edu.sa

Abstract— Image processing and analyzing images in the medical field is very important, this research diagnoses and describes developing of diseases at an earlier stage, a detection of diseases types by using microscopic images of blood samples. Analyzing through images changing is very important, the main objective is completed by analyzing evolutionary computation into its component parts, using elitism immigrants multiple objectives of genetic algorithms (EIMOGAs), artificial intelligence system, evolution methodologies and strategies, evolutionary algorithm. EIOMGAs are the type of Soft Computing a model of machine intelligence to derive its behavior from the processes of evolution in nature [1].

The goal of applying EIOMGAs is to enhance the quality of the images by applying the image converting process segmentation to get the best image quality to be very easy to analyze the images. EIOMGAs are the unbiased estimator for optimization technique, and more effective in image segmentation, and it is the powerful optimization technique especially in a large solution space to implement enhancement process. The powerful of EIOMGAs system in image processing and other fields leads to increase popularity and increasingly in different areas of images processing and analyzing for solving the complex problems. The main task of EIOMGAs is to enhance the quality of the image and get required image recognition to achieve better results, faster processing and implement a specialized system to introduce different approaches based on GAs with image processing to obtain good quality and natural contrast of images [2].The development with comparisons used between the different techniques of representation and fitness analysis, mutation, recombination, and selection, evolutionary computation is shown to be an optimization search tools. All features of microscopic samples images and examines change in geometry, texture, colors and statistical analysis will be applied and implemented in this system.

Index Terms— Elitism Immigrants Multiple Objective, Microarray Image Processing, Data Mining, Digital image processing.

I. Introduction

Image processing is a section of artificial intelligence concerned with the enhancement, and analysis of images performed by a computer, and it has become the most important visualization and interpretation methods in biology and medical fields. It has a development of new and powerful tools for analyzing, detecting, transmitting, storing, and displaying medical images, the medical images is challenging to found the development integrated systems, design, implementation, and successful testing of complex medical systems using in the medical aspect, the analyzing process through images is to collect information, diagnosis diseases, diseases detection, and control and therapy evaluation [3].The segmentation and morphological techniques of Digital image processing (DIP) can be applied for analyzing and diagnosis a lot of medical images diseases such as WBCs, the white blood cells play the main goal in the diagnosis and analysis different diseases, the extracting information is very important for hematologists. The different techniques in an image processing are used to analyze the cells to be more accurate and diagnosis systems for remote diseases. There are some complications to extract some data from RBCs in the cells wide variety in shape, edge, position, and size. Moreover, when the illumination is imbalanced, the image contrast between cell boundaries and the background varies based on the capturing process conditions [4].

In the last few years, the image processing techniques got rapidly grown, where hematologists can be used images segmentations of blood automatically, blood slides and blood boundaries for detecting diseases in the diagnosis system.

The research study is focusing on RBCs segmentation process for human blood system using elitism immigrants multiple-objective of Genetic Algorithms and digital images processing. The main goal is to analyze RBCs using EIMOGAs that has been developed in the last years. The using of EIMOGAs in the segmentation techniques of the digital image processing can be applied set of constraints to finding data about the ratio cytoplasm to classify and identify various types of cells such as a lymphocyte, basophil, and neutrophil. The segmentation methods have been applied in many works and different area of images processing, related to region growing, border detection, edge, watershed clustering, and mathematical morphology and filtering processes.

The author proposed an automatic medical system for the segmentation technique and border identification for whole objects based on image boundary among the images database system that is taken from a blood slides and the original image [5], the using of images processing are used as they are not expensive and do not require complex testing and labs equipment, the system focus on

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Thalassemia disease, Thalassemia features in microscopic images and changes in gene geometry, texture, statistical analysis, and colors contrast of RBCs, therefore, the microarray technology can be applied to get a robust genomic system for studying and analyzing the thousands behavior of genes simultaneously. The images analysis which was obtained from the microarray technology strongly helped in the diagnosis, detection, and treatment of most diseases.

In this research can be developed an automated diagnosis system for analyzing and testing data from microscope images directly and detects diseases cases, for that purpose, the digital image processing performs many operations such as modify image rotation, extracting data from the image, locating genes in the images, and the data mining will be normalized the extracted data and getting the effective genes [6].

I. Genetic Algorithm

Charles Darwin is invented Genetic Algorithm as the natural selection process to take input and calculates an output when a set of solutions can be produced. In the last few years, GAs was created to represent processes in the natural system that is important to evaluate and perform an efficient search in the global domain and to have many optimal solutions and more than that. GA is very effective in the contrast improvement in quality and produces an image based on the natural contrast in different scale levels. GAs are the systematic random search techniques to apply generic methods for solving complex problems and optimization process.In the image process, GA can use less information related to the segmentation problems to be solved than the traditional optimization systems, which almost require the derivative objective functions. The fitness function is based on an individual of images, and additionally, GAs can be used a set of different operators (reproduction, crossover, and mutation) to generate new solutions and use it to get an optimal solution for the new images that may contain new chromosomes [7]. Basically, the new children or chromosomes in Genetic Algorithm are obtained of a combination of features of their parents from original images.

The elitism based immigrants multiple objectives of Genetic Algorithms (EIMOGAs) is a new technique will be used in the image processing to produce a set of newly enhanced pixels of the image to be much better than the original image and contains good features, the Image segmentation will be applied EIMOGAs techniques to enhance and improve image quality for extraction more details about the degraded images. The techniques of image colors have some problems such as colors image enhancement applied in the true colour (RGB), where the colour spaces are not suitable for the human system, and the distribution colours in the images are inappropriate the normal visual limits to human perception[8], one technique is not enough to be suitable for one type of image degradations in the RBCs. EIMOGAs have the ability to select optimal colors and segmentation regions to choose appropriate features of the analysis size and select the heuristic thresholds to solve complex problems [7].

II. Thalassemia diseases

There are some main factors will be used in this research to analyze blood color of RBCs, the cells shape, and the cells number, the experiments diagnosis will be checked whether the required factors are negative or positive results, a lot of diseases can happen to cause the size changing cell, shape cell, and the blood cell color. The researchers can be used blood count analyses, blood images analyze, iron analyzes, and the HPLC analyze to check whether the patients are having thalassemia diseases or not. In this research, we proposed a system that can be applied to diagnosis thalassemia disease based on EIMOGAs techniques, The purpose in this work is to help both patients and doctors and health care regarding the reducing time for pathology, the reducing effort, and more accurate in achieving outputs. In this research will be studied two types of thalassemia disease are alpha thalassemia and beta thalassemia. Thalassemia diseases cause a reduction in the lifespan of red blood cells , the disease is a result of an imperfection in the genes that regulate the haemoglobin formation, which is a core ingredient of the red blood cells ,hence thalassemia is hereditary blood disorder characterized by abnormal haemoglobin production and very common in subtropical and tropical areas, for instance, thalassemia disease was infected 280,000,000 people in 2013, with about 439,000 having a dangerous disease, the most common among Middle Eastern people, African descent, Italian, Greek, and South Asian ,both females and Males have similar disease rates, the resulted in 16,800 deaths in 2016, down from 35 thousand, deaths in 1990, so the blood characteristics should be analyzed to make a good diagnosis.

The automated diagnostics system have been developed, using available rule-based tools to cover a blood broader range related diseases containing anemia various types, the alternative automated diagnostic tools are required, in order to find the diagnostic goal, and the differentiation among thalassemia patients, thalassemia traits, and normal people.

The classification problems of thalassemia patients will be formulated in the pattern recognition problems as input process [9]. The test patterns and samples will be blood-related features that are the red blood cell, characteristics reticulocyte, and blood platelet, that is extracted and used in the blood samples.

In the data mining techniques, the researchers are used different rules and patterns to extract data based on the clustering, summarization, association, and classification using the machine learning techniques to test Beta Thalassemia [10]. There are research studies illustrated the Thalassemia testing indicators as Haemoglobin (Hb) A2, Mean Corpuscular Haemoglobin, and Mean Corpuscular Volume. In the Knowledge research, the principal components analyses research were used to discover β-Thalassemia, there are several algorithms for machine learning are applied in the β-thalassemia

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

classification based on new data set which is different from the other researchers, the classifiers of data mining are applied to differentiate among thalassemia traits in different levels as iron deficiency patients, normal people and the patients with other blood diseases [10].

III. IMAGE PROCESSING AND SEGMENTATION

The image segmentation is the partitioning operation of an image into a collection of pixels connected sets, and it also the most significant task in image processing, and for better analysis and diagnosis, the original image will be partitioned into different sizes and pieces. The most important task in the image segmentation is to explore the appropriate parameter selection based on GAs.The purpose of image segmentation are:

1. The regions segmentation to cover the image coordinates.
2. The linear structures segmentations that including line segments and curve segments.
3. The 2D shapes segmentations, such as ellipses, circles, and strips (regions, long, symmetric) for instance, the cluster pixels inside salient image boundaries, the regions corresponding to objects surfaces, or objects natural parts.

The applications of image segmentation include: the image recognition segmentation is using for face recognition, the medical image segmentation such as diagnosis operations, locating cancer diseases and other dangers pathologies. The image segmentation process was used in the agricultural imaging for crop diseases detection. Traffic control system was used to identify shape and size of objects, and moreover, it used to identify moving scene objects using video compression system, the Image segmentation has been divided into two parts of approaches: the region based approach and boundary based approach, in the first part, the purpose is to determine if a pixel belongs to an object or not[11], in the second part the goal is to locate the boundary curves between the background and the objects.

There are four different types of image segmentation:

a) Segmentation greyscale. b) Segmentation texture.
c) Segmentation motion. d) Segmentation depth.

The Main algorithms of region segmentation are divided into three categories:

1. Region-based segmentation technique:

Thresholding method can be used as a simple technique to segment an image for the objects separating from the background using a pixels features values that are compared with a threshold values in order to determine the class of the pixels, this method starts with the first one pixel of a potential region and expands it by inserting adjacent pixels for any image includes different regions, the image will be segmented based on the different areas of the image which each piece has a range of features values, the thresholds are significant to select these thresholds, and it very effectively and useful in the segmentation quality of the images, finally, the statistical test processes used to take a decision which pixels will be inserted into a region segmentation or not.

2. Clustering-based image segmentation technique is dividing the image into different classes which do not require prior information. In the same type of classes, the data should be collected together in similar classes and the data which contains a different type of classes will be in different classes as possible.
3. Edge-based image segmentation Technique is the main features of the original image, which include valuable data useful in image analysis and diagnosis of object classification and explores the detection of boundaries among the different region's image [12]. The boundaries discontinuities occur among the pixels of the selected features such as intensities, textures, and colors.

IV. IMAGE SEGMENTATION USING GENETIC ALGORITHM.

The parameter selection will be applied using EIMOGAs to enhance the parameters selection of the images segmentation and to improve its outputs. The pixel scale and level of segmentation implement GAs will be used to complete region labeling tasks of the image segmentation processes, the proposed method should be used the image adaptive segmentation including the following steps [10]:

1. Compute the image statistics tables give us the probability for a given degrees of a confidence level and identically distributed normally to select suitable threshold. 2. Generate an initial population of segmentation image. 3. The image Segmentation applied initial parameters selections. 4. Compute the segmentation based on quality measures to satisfy conditions of the fitness function.
5. To select new individuals should be used the reproduction operator to generate new population applying by using the mutation and crossover operators.
6. The image segment should be used new parameters to calculate the segmentation quality of an image.
7. Analysis and modify the knowledge based on the knowledge structures of the new image.

V. GENETIC ALGORITHM AND CHROMATIC FEATURES

In this research, we will be applied Elitism Immigrants Multiple Objective of Genetic Algorithms (EIMOGAs) with Chromatic features to describe the color distribution and the grey-level of the images, which are the most discriminative features of Red Blood Cells. the image pixels is represented a segmented object such as (RBCs, RBCs, Nucleus, Cytoplasm, Cells Parasites), The GAs selection operator is used to detect the edge of cells boundaries that have the same colors of pixels from the current population (RBCs images) that will be used new generation.

The convergence process will be completed and achieved in under the iterations required the number to detect RBCs and complete blood counted for a new generation and population.

In the next step of the population, solutions are represented intensity cells colors and chromatic features which can be detected and computed using EIMOGAs of RBCs. In this stage of research will use generation,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

mutation, selection, elitism immigrants, integrate all several immigrants of memory scheme and combined into the EIMOGAs to improve its searching capacity for the image process environment.The image process is a stochastic process where pixels values are modeled as random variables, the GAs can be applied to calculate the probability density of grey level and color distribution as its fitness function [13].

The fitness function is used to get robust convergence as building simulations for RBCs image as possible with reliable convergence and a high convergence with the original image.

the elitism-based immigrants schemes Multi-objective of genetic algorithms (EIMOGAs) efficiently improve the GAs performance in the image processing environment ,and the best selection individual of color pixels based on fitness function from the previous generation is used to create immigrants included probability density of grey level , colors gradient ,colors distribution, cells color and boundaries shapes of RBCs into the population using a genetic operation (selection, evaluation, mutation ,and recombination) , the new process of generation will be the implemented using the elite process e_t from the previous generation g_{t-1} to create new immigrants, as a set of r(e) × n(t) individuals are generated based on fitness function and mutation e_{t-1} with a probability $p(e_{t-1})$, where n(t) is the population number of the image colors , and r(e) is the number ratio of elitism immigrants for each color to the population number, the selection operator of EIOMGAs selects set of cells color of RBCs as the best solutions that have a better classification, based on a fitness functions [14], and then it will be carried forward for recombination image process.

The sensitivity analysis and the results are shown in the final experiments. EIOMGAs are efficiently improve the genetic algorithms performance in the image processing environment, and the best individual from the previous generation (RBCs Image) to next generation can be selected and created immigrants with optimal solutions into the population by evaluation and mutation process.

VI. Proposed System

The proposed analysis system for RBCs segmentation explants the phases are shown in Figure 1. The image pre-processing of the blood smears is applied for removing noises, improving and contrast variation and luminance in the original images. In the second phase, a segmentation processes are applied and implemented to explore and isolate the interest objects of the image. The third phase goal is to extract the objects characters to be used in the next phase of the process, the Features selection method is applied to decrease the redundant data and built classification stage. The selected features are selected for input to the classification method and take the decision about the class assignment by using EIOMGAs as shown in figure 1.

The main goal of the segmentation process is to separate RBCs from another different ingredient of blood image. The blood smear consists four components, the image background, WBCs, RBCs, and cytoplasm. WBC should be darker than the background, and RBCs seem on a high-intensity scale [15]. And also, there are shapes variation in cells and their nucleus.

Figure 1 shows the block diagram of the segmentation scheme.

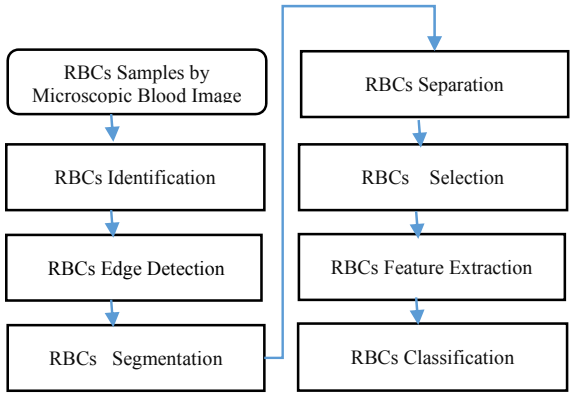


Fig.1.The Proposed Block Diagram of RBCs Analysis and Methods Using EIMOGA

VII. Image Segmentation and Active Contour Models

In the last few years, there are recent developments in the medical imaging fields have brought a new techniques research on image processing for improving medical analysis and diagnosis in segmented images. This technique has been developed to identify specific structures in a magnetic resonance imaging (MRI). The Active Contour methods are adaptable to the desired features in the image.

There are several forms and different types of RBCs images. The applying appropriate method for variable shapes and segmenting for RBCs has been always a challenge for researchers between segmentation methods, the active contour model has a lot of enhancements and implemented in the last few years, In the RBCs, the image should be used active contour models which are changeable curves to respond their change forms to avoid deform objects boundaries in an image segmentation [16].

The active contour models can be moved based on internal or external forces extracted from the image characteristics. The active contour adaptation occurs in response to both internal and external forces, the external forces model has described the gray level gradient, the active contour models can be divided into two types: the parametric models like the Snakes model, which defines a resilient contour that can dynamically adapt to required edges of the image objects, and the geometric models, such as the Level Set model [13] it embeds the front to be zero level set in the higher dimensional function, to calculate the new function evolution, this evolution operation is dependent on the image characteristics extracted and geometric restrictions of the function.

In the processing scheme, the segmentations are implemented on sub-images, the parametric snake model is a curve x(s) defined in Equation.1 [4], to move through

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the image spatial domain and minimize the energy function E(s) defined in Eq.2.

$$v(s) = [x(s) \ , \ y(s)] \ , \quad s \in \{1,0\} \qquad (1)$$

$$E(s) = \int_0^1 \left[\frac{1}{2}(\alpha \mid x'(s) \mid^2) + \beta \left| x''(s) \right|^2 + Ex(x(s)) \right] ds \ (2)$$

Where x'(s) denotes the first derivative, x″(s) denotes the second derivative of x(s). While α, β are parameters of weighting to control the rigidity and tension of snake, respectively. E_{xt} is the function of external energy which is derived from the image to take smaller values of features of boundaries.

The external and internal forces are used the image gradients as a parametric active contour of Snakes models, the external and internal forces will be used the image gradients as a parametric active contour of Snack models. The gradient-based model is better models to use insensitivity to its initial parameters positions and wide capture region of images. Gradient vector flow detriments the object boundaries when addressed locked to the object boundary, while on homogeneous regions will be changed smoothly and will be more extended to the image border, the gradient vector flow field is selected as v(x, y) to be vector field is written in equation.3 [17], which reduces the energy function as defined in the equation. 4.

$$v(x,y) = v(x,y) \ , \ u(x, \ y) \qquad (3)$$

$$E(s_{min}) = \int_0^1 [(\alpha \mid x'(s) \mid^2) + \beta \mid x''(s) \mid^2 + E(x(s))] ds \qquad (4)$$

The internal energy E for the gray-level images I (x,y) is identified as:

$$E_{int} = \int_0^1 \frac{1}{2} [(\alpha \mid x'(s) \mid^2) + \beta \mid x''(s) \mid^2] ds \ (5)$$

While the external energy can be identified as:

$$E_{ext} = -\nabla [(g(x, \ y) * I(x, \ y) \mid^2] \qquad (6)$$

Where g denotes a two dimensional of the Gaussian filter with a normal deviation, ∇ is identified the gradient operator. This filter is applied to the image in order to enhance the map image edge and to reduce an image noise . The regions closer to edges will be given the gradient image high rates. In this research, the cell boundaries will be extracted using edge detection and avoided missing off the edges occurring[18], so the image smoothing using a Gaussian filter to reduce noises with normal deviation is written as the following [17].

$$g(x,y) = G_\delta(x, \ y) * f(x, \ y) \qquad (7)$$

Gaussian's smoothing operator (GSO) is a 2-Dim used to remove noise and detail with special properties as defined in Eq.8.

$$G(x,y) = \frac{1}{2\pi\delta^2} e^{-\frac{x^2+y^2}{2\delta^2}} \qquad (8)$$

There is another problem needed to a solution using additional parameters to improve the external force (k, k1) to improve the capture range in heterogeneous regions of the image edges. This problem will be enhanced using a constant normal (k) [19], to control the external force, the active contour model can be inflated or deflated, based on the sign and magnitude (k1) of the external force. in this paper, the author is proposed applying the balloon model to prevent the snake from stalling in the image homogeneous regions and should be taken to select appropriate values to (k, k1) to make the snake control edges and noise, without exceeding the desired characteristics for contour regions , which is written in Eqau.9.

$$F_{ex} = F_1 \vec{n}(s) - k \frac{E_{ext}}{\|E_{ext}\|}(x(s), y(s)) \qquad (9)$$

Energy Surface and Optimum Thresholding is the basics approach to image segmentation is an amplitude thresholding, a threshold T is chosen to separate the two regions modes, the image point for I(x,y) >T is considered as object points[19], otherwise, the point is called a background point. The threshold method is defined as:

$$g(x,y) = \begin{cases} 0 \ , & I(x,y) < T \\ 1 \ , & I(x,y) \geq T \end{cases} \qquad (10)$$

Where T is set on the entire image basis I(x, y), and the threshold is global. When T depends on spatial coordinates x and y, based on a dynamic threshold, when T depends on both I(x,y) and set property p(x,y) of local image, the average of gray level in a neighborhood centered on I(x,y), the threshold will be local and T is set according to a fitness function is defined by:

$$f(y,x) = \quad T[p(x,y), I(x,y)] \qquad (11)$$

Template Matching is a new type technique in the image segmentation based on prior knowledge of the detected object in image analysis, using the presence detection of an object in a scene, and identifies its position in current given scene [20].

The object locating can be described using a template T[x,y], in the image I[x,y], The best match Searching can be minimized the mean squared errors as written below:

$$E[p,q] = \sum_{x=-\infty} \sum_{y=-\infty} [I[x,y] - T[(x-p),(y-q)]]^2 \ (12)$$

In this research, the correlation technique of images will be used for exploring match of a searched shape w(x, y), of size k*l within an image I(x, y)of a larger size m × n. Where w(x, y) is a search of shape, shape size denotes as z (l, k) , maximum size as m(l, k), the summation will be taken in the image region when w and I not separately, the correlation function techniques have the sensitive disadvantage to local intensities of w(x,y) and I(x,y), the correlation coefficient C(s,t) can be used to remove difficulty pattern matching of local intensities as follows.

$$C(s,t) = \frac{\sum_x \sum_y [I(x,y) - \bar{I}(x,y)] \ [w(x-s,y-t) - \bar{w}]}{\sqrt{\{\sum_x \sum_y \ [I(x,y) - \bar{I}(x,y)]^2 \ \sum_x \sum_y [w(x-s,y-t) - \bar{w})]^2\}}} \qquad (13)$$

In the image, analysis can be used Hough transform technology as a technique of a feature extraction image for RBCs number and to get the number of red blood cell count in the image. Then using machine learning algorithms tool, which has developed a formula to convert a number of red blood cells in the image to actual count by Hough Transform, blood count calculates the blood cells number in a cubic millimeter of blood volume. In this research can be calculated the number of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

RBCs per cubic millimeter based on the cells number in the given image [20].

RBCs count per_cumm = RBCs / ((S /M^2*) * dilution_factor.

Where (cu, mm) is cubic millimeter, and (µl, mcl) is microliters, them is magnification.T is a film thickness. S is an image size.

The Circular Hough Transform technique can be used as a measuring tool to calculate the accuracy using the result of Red Blood Cells number compared with manual counting as the following:

$$\text{Accuracy} = \left(\frac{\text{RBCcount}}{\text{ActualCount}} \right) * 100\% \quad (14)$$

The RBCs classification results need to use set of the numerical analyzing as parameters: mean corpuscular volume (MCV), RBCs distribution width, RBCs count, mean corpuscular hemoglobin (MCH), hemoglobin count (HB), and mean corpuscular hemoglobin concentration (MCHC), for identifying the analyzing combinations.

The hemoglobin (HB) is responsible for the red blood cells color, HB Contents can be calculated by measuring the gradient of colors, the threshold Technique an image will partition into two parts: the foreground and the background.

The binary algorithm was used to calculate the values colors in an image. The classification tools used to formulate a formula and calculate the hemoglobin contents[21][22].

There are three parameters can be identified the red cells characteristics including reporting units, formulas, and definition to calculate each parameter as shown below.

MCV is the RBCs average size constituting the sample. One femtoliter is 10-15 L, the adult's interval (80 - 100 fL) is defined by.

$$\text{MCV} = \left(\frac{\text{Hematocrit} * 10}{\text{RBC} * (10^{12}/L)} \right) * 100\% \quad (15)$$

MCH is the hemoglobin average weight in the RBCs, one pictogram is 10 -12 grams, adults interval (26-32 pg) is defined by

$$\text{MCH} = \left(\frac{\text{Hb (g/gL)} * 10}{\text{RBC} * (10^{12}/L)} \right) \quad (16)$$

MCHC is the hemoglobin average concentration in the RBCs, adults interval (32 - 36 g/dL) is defined by.

$$\text{MCHC} = \left(\frac{\text{Hb (g/gL)} * 100}{\text{Hematocrit}} \right) * 100 \quad (17)$$

The blood samples need to be diluted, so there are some of the RBCs in whole blood to be accurately counted in a microscope[23]. The dilution factor (*Df*) is provided by the industrialist and is around 200X as shown below in equation 18.

$$\text{Actual RBCs} = \frac{\text{RBCs counted}}{(\frac{Ia}{mf^2}) * \text{Ft}) * \text{Df}} \quad (18)$$

VIII. Results and Discussion

This section will be assessed the performance of the proposed RBCs segmentation system. In our experiment, 22 blood samples images of thalassemia. The Image captured was digitized using a digital video camera of Sony high resolution, which was coupled to a Microscope LCD Biological BX5.The experiments were implemented using the Java Genetic Algorithms Package In this research study, there are many methods to diagnosis the Thalassemia diseases. There are three types of images using proposed method to discover abnormal cell types, the color red of blood cells and classification tools.

The research results are described in the following sections:

1- The researcher has identified the abnormal cell types using the various images of red blood cells. The researcher experiments have classified the red blood cells in different shapes. Figure.1 shows the different shapes of red blood cells and the changes in the colors rates (Red, Green, and Blue) of the images after the calculations process have done, the analysis explained the relationship between Thalassemia and non-thalassemia blood images that it has taken as shown in Fig 1.

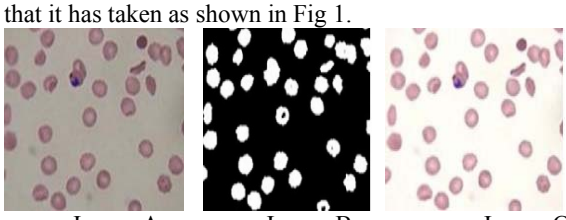


Image A Image B Image C

Fig.1. shows the results of thalassemia infected.

In these experiments , there are different results of Image processing for thalassemia infected, image A is the original image, Image B explained blood image, and Image C shows results after applying pixel classification using EIMOGAs as shown in figure1.

The target cell of blood image, which has hypochromic microcytic and abnormal cell such as sickle cell, that patient will be an affected person of a thalassemia, and for further classification, the author uses other extracted information from classification tools.

2. The RBCs calculations are made to extract the color information (Red, Green, and Blue values) for each image using image processing. In this research, the studies have calculated the average intensity for each of (red, green, blue color), that gave the average rate of colors (red, green, blue) for each sample blood image. The results of average intensity identified the R, G and B rates of thalassemia normal image rate >185 and thalassemia abnormal blood image rate <= 185.

The blood samples need to be diluted, so there are some of the RBCs in whole blood to be accurately counted in a microscope. The dilution factor (*Df*) is provided by the industrialist and is around 200X as shown below in equation 18. The results are reported as the number of RBCs per cubic millimeter of blood.

Males Normal Values = 4.2 - 5.4 million RBCs / mm 3, Females = 3.6 - 5.0 million RBCs / mm 3.

Stained thin blood should be taken by Digital microscope to be more easily distinguished between platelets, RBCs, and WBCs.

To differentiate between RBCs, WBCs, and Platelets, RBCs is less stained as compare to WBC and platelets leaving a bright spot and its intensity value similar the background value.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1. Standard complete blood count System for healthy people.

Blood cell type	Women	Men
RBCs	4-5 M/µL	4.5 - 6.0 M/µL
WBCs	4.5 - 11 K/µL	4.5 – 11 K/µL
Platelets	150 – 450 K/µL	150 - 450 K/µL
Haematocrit	36% to 45%	42% to 50%
Haemoglobin	12 - 15 gm/100 ml	14 - 17 gm/100 ml
Gm /100ml: gram per 100 milliliters; ml: milliliter; gm: grams; µ: Microliter; K: Thousand; M: Million.		

After isolating process of RBCs, we need to apply a counter for counting the RBC number in the image process, so we have used a formula to calculate RBCs number per cumm on the cells number in the given image area (*Ia*) of the blood samples and the of the blood sample film thickness (Ft) is 0.1 mm that is the standard medical system. The magnification factor (Mf) which is the magnification level under the microscope.

In the three experiments below, the assessing RBC morphology procedure includes the smear examination in the thinner edge where the erythrocytes (RBCs) are randomly distributed, the most part singly, with sometimes overlapping cells.

In the next experiment, figure 2 shows three images, Image A is shown an acceptable area of RBC morphology evaluation, we can see that the most cells clearly can be distinguished with some overlapping cells. In an image B is included area too thin, the RBCs appear very flat such as shape cobblestone. An Image C is shown the examined area is thicker than Image B, so the cells will close together, the evaluation process of the morphology will be used individual cells.

Image A **Image B** **Image C**

Fig 2. Segmentation thinner edge area.

The experiments results of the smear examination in the thinner edge area and Assessing Erythrocytes (RBCs) Morphology as shown in figure 2.

Table 2. The Result Comparison between Counted Manually System and CHT using EIMOGA in an Image Processing.

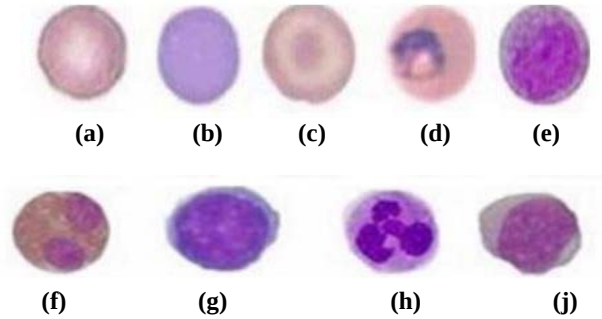
Images	Radius Range In pixels	Film Thickness	RBCs counted Manually System	RBCs counted proposed System
Img1	5-12	0.1	3.768	3.542
Img2	5-14	0.1	4.349	4.161
Img3	4-15	0.1	6.821	6.742
Img4	4-14	0.1	5.783	5.621
Img5	5-13	0.1	4.981	4.813
Where: Magnification Factor(Mf)=300*300; Dilution Factor(Df) =200 ;				

Table 1. Type Sizes for Camera-Ready Papers

After the calculations that have done. In Fig 5 shows the changes in the colors rates (Red, Green, and Blue) of the two Images.The analysis explained the relationship between Thalassemia and non-thalassemia blood images that it has taken.

In the research observations, based on the research details and results, could able to analysis possible diseases combinations when high or low above according to values of parameters.

In figure 3, shows if there is a normal or abnormal of MCV, MCH, RBC and MCHC attributes of a patient's blood sample, and this analysis can be diagnosed the current disease for that patient.



(a) **(b)** **(c)** **(d)** **(e)**

(f) **(g)** **(h)** **(j)**

Fig 3. Different Images of RBCs samples using EIMOGA

Figure.3 shows different Images of RBCs samples using EIMOGA as the following: the images (a) and (b) are healthy RBCs. the images (c) and (d) are infected RBCs with ring parasites. The Images (e) and (f) are infected with trophozoites parasites. The images (g), (h) (j) are RBCs infected with schizont parasites.

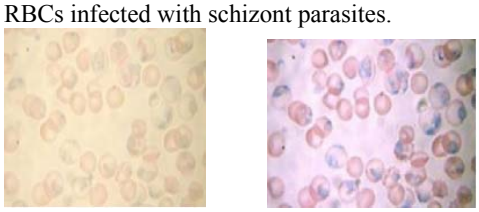


Image A **Image B**

Fig4. Compression between original and enhanced image.

In figure 4, Image A corresponds to original Image, and Image B displays enhanced image after applying EIMOGAs approach for image filtering.

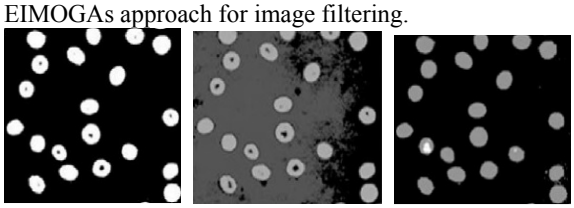


Image C **Image D** **Image E**

Fig 6. Shows different method segmentation results.

Figure 6 shows three different method segmentation results of images. Image C is gotten using EIMOGAs algorithm of the histograms approach. Image D is obtained by applying EIMOGA for pixels classification. Image E is used EIMOGA for clustering approach.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

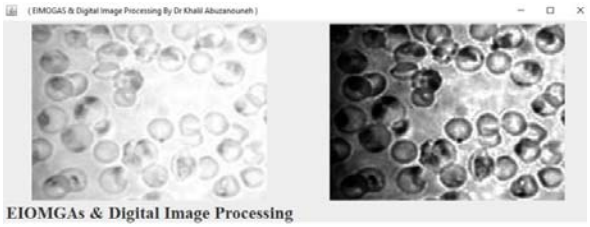


Image A Image B

Fig 7.Shows EIOMGAs technique to enhance Images results.

The experiments results using EIOMGAs technique to enhance the contrast of images,Image A is gotten using EIMOGAs algorithm of the histograms approach. Image B is obtained by applying EIMOGA for pixels classification as shown in figure 7.

In this experiment as shown in table 3, the Original Image number is 5 of microscopic image, the accuracy average is 93.33% was implemented. The comparing process is implemented between the input image and an image after using Circular Hough Transform, there are some of RBCs is not calculated in CHT method due to deformable shape and another condition.

Table 3. Shows the RBC results are counted for 5 RBCs Images using CHT method.

Original Image	Actual RBCs	CHT method Counted	Accuracy of CHT
Img_RBCs_1	903	837	92.69
Img_RBCs_2	937	871	92.96
Img_RBCs_3	978	916	93.66
Img_RBCs_4	1018	957	94.01
Img_RBCs_5	1123	1048	93.32
			Average= 93.33

In the next experiment, we have the same number of Original Images from the previous experiment microscopic image, the accuracy average is 97.05 was implemented. The comparing process is implemented between the input image and an image after using EIMOGA to enhance analysis process, the final results were more accuracy for the counted process of RBCs by EIOMGA as shown in table 4.

Table 4. Shows the Accuracy Results for 5 RBCs Images using EIMOGAs method.

Original Image	Actual RBCs	EIMOGAs Counted RBCs	Accuracy of EIMOGA
Img_RBCs_1	903	877	97.12
Img_RBCs_2	937	913	97.44
Img_RBCs_3	978	949	97.03
Img_RBCs_4	1018	987	96.95
Img_RBCs_5	1123	1086	96.71
			Average= 97.05

IX. REFERENCES

[1] Khalil Ibrahim Mohammad Abuzanouneh. “Hybrid Multi Objectives Genetic Algorithms and Immigrants Scheme for Dynamic Routing Problems in Mobile Networks”. International Journal of Computer Applications 164(5): 49 -57, April 2017.

[2] Khalil Ibrahim Mohammad Abuzanouneh “Parallel and Distributed Genetic Algorithm Multiple Objectives to Improve and Develop of Evolutionary Algorithm, International Journal of Advanced Computer Science Applications, Volume7 Issue 5, 2016.

[3] Ritter Nand CooperJ(2007)Segmentation and border identification of cells in images of peripheral blood smear slides. In Proceedings of the Thirtieth Australasian Conference on Computer Science 62:161–169.

[4] Gonzalez RC, Woods RE, Eddins SL (2003) Digital image processing using MATLAB. Prentice-Hall, Upper Saddle River.

[5] Ritter Nand CooperJ(2007)Segmentation and border identification of cells in images of peripheral blood smear slides. In Proceedings of the Thirtieth Australasian Conference on Computer Science 62:161–169.

[6]

[7] Cseke I (1992) A fast segmentation scheme for white blood cell images. In Pattern Recognition. Conference C: Image, Speech and Signal Analysis, Proceedings. 11th IAPR International Conference on vol. III:530–533.

[8] SobelI(1978) Neighborhood coding of binary images for fast contour following and general array binary processing. Comput Graph Image Process 8:127–135.

[9] Prewitt JMS (1970) Object enhancement and extraction. Picture Processing and Psychopictorics, pp. 75–149.

[10] NHLBI. "How Can Thalassemias Be Prevented?". 3 July 2012. Archived from the original on 16 September 2016. Retrieved 5 September 2016.

[11] ZackG, RogersW, Latt S(1977)Automatic measurement of sister chromatid exchange frequency. J Histochem Cytochem 25:741–753.

[12] Otsu N (1975) A threshold selection method from gray-level histograms. Automatica 11:285–296.

[13] Ongun G, Halici U, Leblebicioglu K, Atalay V, Beksac M and Beksac S (2001) Feature extraction and classification of blood cells for an automated differential blood count system. Neural Networks. Proceedings. IJCNN '01. International Joint Conference on vol. 4:2461–2466.

[14] Kan Jiang, Qing-Min Liao and Sheng-Yang Dai (2003) A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering. Machine Learning and Cybernetics, 2003 International Conference on, vol. 5:2820–2825.

[15] Dorini LB, Minetto R and Leite NJ (2007) White blood cell segmentation using morphological operators and scale-space analysis. In SIBGRAPI '07: Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing.

[16] Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. Comput Graph Image Process 1:321–331

[17] Canny J (1986) A computational approach to edge detection. IEEE Trans. Pattern Anal Mach Intell, pp. 679–698.
[18] Xu C, Prince JL (1998) Snakes, shapes, and gradient vector flow. IEEE Trans Image Process 7:359–369.
[19] Xu C and Prince J (1997) Gradient vector flow: a new external force for snakes. In 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997. Proceedings, pp. 66–71.
[20] Kumar BR, Joseph DK, and Sreenivasc TV (2002) Teager energy based blood cell segmentation. Digital Signal Processing. DSP 2002 14th International Conference on, vol. 2:619–622.

Dr. Khalil Ibrahim Mohammed Abuzanouneh is assistant professor in Computer Collage at the Department of Information Technology, Qassim University. In teaching, he has been focusing on applying genetic algorithms and solves problem based on genetics approaches in Computer Education. In research, his current interests include optimization problem and data compression and security. Dr. Khalil received his PhD degree in Computer Engineering Ph.D.in Computer Engineering Sciences “Computer Aided Design”.
.

Reliability of Scalable Coherent Interface in Cluster Interconnection

Mr Adinew Belay[1], Dr Narasimha Rao Yamarthi[2]

HOD, Department of CS, SCI, Mizan Tepi University, Eithiopia.
Professor, Department of CS, SCI, Mizan Tepi University, Eithiopia.
adinewb2@gmail.com[1],narasimha.yamarthi@gmail.com[2]

Abstract: In recent advanced technology parallel computing plays a vital role in High performance computing. The processor interconnection is one of the prominent factor decides the enactment of high performance computing. The clustering of processors in parallel computing is a challenging job for the present engineers as it depends on many parameters to monitor in packet communication. In the present paper awell-organized scalable coherent interconnection is presented to improve the efficiency of comprehensive parallel computing and cluster reliability. The reliability analysis of redundant fault-tolerance systems is discussed with Markov Modelling. This can be greatly helps to design an efficient and fault tolerance cluster interfacings. The Symmetric clustered processors are designed using Proteus simulation tool.
Keywords: Cluster Interconnection, Markov model, Coherent Interface, Reliability, Scalable system

1. Introduction

Multiple processors used in a computer system enhance the speed of the user operations. Adding second processor is very easy and economical when compared with the second additional computer. Multiple CPUs are interconnected to meet the above purpose. Multiple CPUs are used to perform multiple tasks in single time. Each CPU dedicated with individual task. Such as one CPU controls the operating system programs, and the other CPU may control either memory or I/O operation. Multiple programs with multiple set of instructions can be executed by individual processor [1]. Multiple CPUs are connected in single computer and they are able to perform multiple operations simultaneously and can simultaneously allocate tasks to individual processes or programs [2][3][4]. Although there are multiple CPUs, there are still some difficulties need to be addressed for reliable operation in the computer system [5].

Multiple processor interconnections are implemented such that each CPU and each interconnection behaves separately with or with-out matching their functionalities. There are few limitations need to be face by adding multiple CPUs or second CPU in the multiprocessor system. In general the Operating System (OS) developed in a system for supporting and configuring single CPU. On the other hand when second CPU is connected, then operating system took log time to configure the second CPU and also takes long time to run the tasks in the second CPU. Hence to handle several CPUs it took long time to make then perform reliable and efficiently. This problem is overcome by extend its minimal services of operating system with the second CPU too. The extending time period help to configure the second time. But, the extended time periods slow down the all tasks, and make it enter into boot process. On the other hand, the OS executes the tasks given at the primary CPU. The alternate process extending the services of the OS may not satisfy the actual requirements of the parallel processing. Indeed it is decreasing the performance and led to diminishing the reliability. On the other hand the common OS allowed to run all processors, but not allowed to run programs on peripherals/IO devices or on particular peripheral or on particular peripheral on particular CPU [6].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In multiprocessing each processing system shares a common main memory and peripherals for simultaneous allocation of tasks [7]. It is not always true that multiple processor uses single task or process simultaneously. The systems which share all CPUs in the same way are called Symmetric Multi-Processing (SMP). The Systems which does not treat all CPUs in the same way and the resources are used in different ways are called Asymmetric Multi-Processing (ASMP), clustering multiprocessing, and Non-Uniform Memory Access (NUMA). In the present work symmetric multiprocessing is implemented using Proteus simulation tool.

The processors are either tightly or loosely coupled in multiprocessor systems. In tightly coupled systems the CPUs are coupled at the bus level and shared with common central memory or in some systems hierarchical memory system is shared. The size of the tightly couple processor is physically small and perform is better than loosely coupled multiprocessor system. The chip type multiprocessor systems and multiprocessors in mainframes are of type tightly coupled. The standalone and single or dual processors are connected through high speed Ethernet in loosely coupled multiple processor systems. Loosely coupled multiprocessors are less expensive when compared with tightly coupled multiprocessor system. Tightly couple systems are efficient in power consumption. Loosely coupled systems can work at different operating systems and versions.

2. Computer Cluster

The computer cluster may contain either tightly coupled or loosely coupled multiprocessor system. The coupled systems in multiprocessor system work together to form a network or to form a single system. Each node in the computer cluster is controlled, scheduled, and managing the tasks by software [8].Each node in the cluster is connected with local area networks. In most of the instances the nodes in the cluster use same type of hardware and OS. In few applications like open source applications different hardware and operating systems are used in cluster nodes [9][10]. The load in the computing system will be distributed to all nodes in the cluster. This will optimize the load queries assigned to the cluster nodes. Distributing of load to optimize the load is called load balancing. The cluster computing is used in simulated weather analysis rather than database analysis. The cluster performance mainly depends on scalability, low maintenance, centralized management, and resource allocation.

3. Cluster Management

One of the major challenges in cluster is management of all processors in the cluster. Sharing of memory in cluster is difficult to manage and cost effective. In a heterogeneous processors cluster the individual tasks are given and the performance of the job is decided by the models and characteristics of the cluster. Hence mapping of various tasks into clusters produce major challenges. When a node in the cluster the method called fencing plays an important role in making the system operative. The fencing can be done in two ways. One is deactivate the particular failed node or disallow others to access the resources [11]. The CPUs in the cluster randomly changes to estimate the future states in the operations allocated to them. The Markov model is the stochastic method used to design randomly changing system, which are depending on present state but not on past state.The Markov random field depends on the neighbouring states in multiple directions. Based on the observations made on system, the systems are divided into two methods. Neighbouring states can be estimated by distributing the random variables. The distribution further depends on neighbouring variables. One is autonomous system, which includes Markov chain model, Hidden Markov model. The second is controlled

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

system which includes Markov decision process and partially observable Markov decision process [12][13].

3.1 Markov Chain: It is a simplest Markov model. Markov model which depends on previous state. It states that the randomvariables vary as the time passes.

3.2 Hidden Markov Model: It is markov model in which states are only partially observable. But those observations are not sufficient to define the state. These sequence of observations are evaluated by forward algorithm and starting probabilities, observation functions, are evaluated by Baum-welch algorithm.

3.3 Markovdecision Process: The transitions in the model depend on current state and the action vector. Reinforcing algorithms are implemented in this process and solved with iteration methods.

3.4 Partially observable Markov decision process: It is a process in which states are only evaluated partially. They are mostly used in artificial intelligence applications such as robotics.

4. Architectural Design

In the present work Symmetric Multiprocessors are interconnected and data transfer is observed using simulation tool Proteus. In symmetric multiprocessor system two or more processors are connected to common system bus and shared a common main memory. All the processors are grant permission to access memory and IO devices. A single OS is used to control all processors and IO systems. The polling of IO devices are controlled by interrupt controller programs. The SMP architectural block diagram is shown in figure 1.

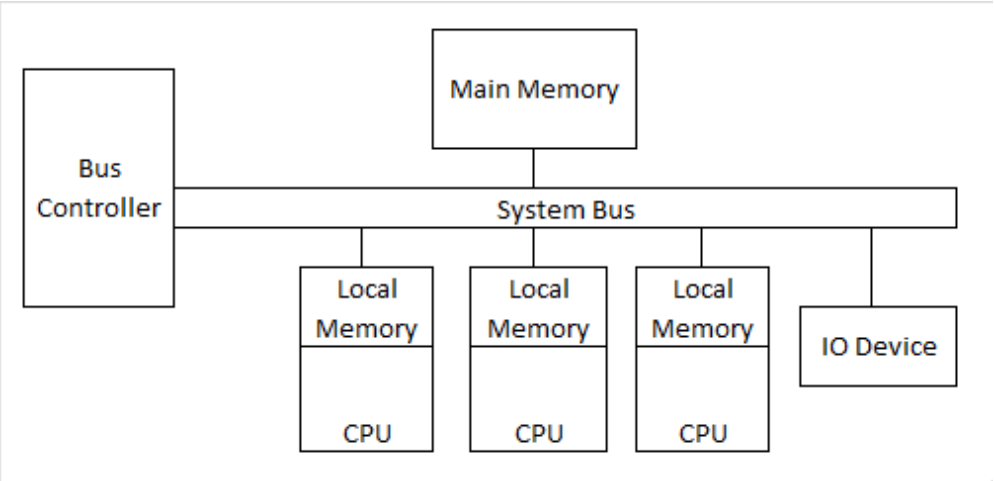


Figure 1: Architecture of Symmetric Multiprocessor System

The bus controller will control the direction of data and produces the control signals like read, write, and interrupt. Memory and IO selections are done by bus controller. All processors have equal priority to access the main memory. The fetching from the memory is controlled through daisy chain method. The memory access is done through system bus. Only one operation either IO transfer or Memory transfer is done with the help of bus controller.

In the present architecture Markov network is considered for analysing the processor operations in multiple dimensions. Unlike Markov chain the Markov network each state depends on neighbours located at different directions. Markov network is visualised like graph of variables.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Here the reliability of the cluster-based system is defined as follows,

$$R_{cl}(t) = R_c(t)\sum_{l=2}^{k} p_l(t) \text{ ---------- (1)}$$

Where,R_c (t) is the reliability of the bus controller, and the probability of *l* number of connected functional clusters with the system at time t is $p_l(t)$is given by,

$$p_l(t) = C_C^{k-1}\binom{k}{l}(R'_C(t))^l(1 - R'_C(t))^{k-1} \text{ ---------- (2)}$$

5. Results

In the current work the throughput is observed in data transfer from cluster to memory. The throughput is observed while all processors in clusters are interconnected without any disconnection between them. The interconnection is framed so that the data successfully transferred to destination even if any processor node failed or disconnected. When any node is failed the data transfer from the failed node to be transferred will be bypassed and the data will be bypassed to next or neighbouring node. Then the data of failed will be passed to destination through neighbouring node along with the new bypassed node data. The table 1 shows the number of active nodes and their respective throughput.

Table 1: Throughput Vs failure nodes

Number of failed Nodes	Throughput (µs)	
	Traditional Model	Markov Model
0	0.825	0.825
1	1.15	1.05
2	1.40	1.25
3	1.65	1.45

6. Conclusion

Hierarchical Markov model is applied in the present work to classify the states of CPUs present in the multiprocessor system. For instance switching the states between memory fetching, IO device controls by multi-processor is observed. Efficient and reliable transfer of data and switching the states between processor is successfully handled. The present analysis is proved that the Markov models are effectively implemented to subvert architectures of advanced cluster-based systems. The reliability is enhanced in terms of throughput and observed in case of failure nodes also.

References

1. Introduction to Multiprocessing (http://www.bitsavers.org/pdf/dec/pdp11/1174/ Introduction _to_Multiprocessing _Jun79.pdf): distinguishes “symmetric” from “master/slave”.
2. Raj Rajagopal, “Introduction to Microsoft Windows NT Cluster Server: Programming andAdministration”,CRC Press, 1999, p. 4.
3. Mike Ebbers; John Kettner; Wayne O'Brien; Bill Ogden, “Introduction to the New Mainframe: z/OS Basics”,2012, IBM. p. 96. ISBN 9780738435343.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

4. Chip multiprocessing (http://www.computerworld.com/hardwaretopics/hardware)
5. Kumar S et al., “Optimizing Technique to Improve the Speed of Data Throughput through ipeline”, Information Systems Design and Intelligent Applications, 8-9 Jan 2016, by SPRINGER Advances in Intelligent Systems and Computing (AISC series), Vol 435.
6. Early Computers at Stanford: the dual processor computer at the AI labhttp:// forum.stanford.edu/wiki/index.php/Early_Computers_at_Stanford#DEC_PDP10_2.
7. Irv Englander, “The architecture of Computer Hardware and Systems Software”, An Information Technology Approach, 2009, 4th ed., Wiley.p. 265.
8. IEEE Technical Committee on Scalable Computing, TCSC, https://www.ieeetcsc.org/.
9. Kumar NS et al., “A New Method to Enhance Performance of Digital Frequency Measurement and Minimize the Clock Skew” in IEEE Sensor Journal, vol 11(6), 2011.
10. Kumar et al., “Design of Intelligent Multinode Sensor Networking”, International Journal on Computer Science and Engineering (IJCSE), vol2 (3), 2010, pp. 468-472.
11. K. Shirahata, et al., “Hybrid Map Task Scheduling for GPUBased Heterogeneous Clusters”, Cloud Computing Technology and Science (CloudCom), 2010 Nov. 30 to 2010 Dec. 3, pp. 733 740.
12. Leslie Pack Kaelbling; Michael L Littman & Anthony R Cassandra, "Planning and acting in partially observable stochastic domains", Artificial Intelligence (Elsevier), 1998, Vol 101, (1–2): pp 99–134.
13. H. H. Bui, S. Venkatesh, and G. West,“Policy recognition in the abstract hidden markov model”, Journal of Artificial Intelligence Research, 2002, vol. 17, pp. 451–499.
14. Yue Luo et al., “Analyzing and Improving Clustering Based Sampling forMicroprocessor Simulation”, Proceedings of the 17th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD’05), 2005, IEEE.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Effort Prediction Tool using UML diagrams (EPT)

[1]Atica M. Altaie, [2]Asma Yassin Hamo

[1] Atica_swe@yahoo.com, [2]Asmahammo@yahoo.com
[1,2] *Software Engineering Department, Computer and Mathematics Science College, University of Mosul, Mosul, Iraq*

***Abstract*-The use case and class diagrams are important models of the system created during early phases of the software development. Effort and size estimation are also important points of the software development. Many effort estimation models proposed in the last years and many factors have an impact on software efforts like complexity, use case points and class points. Effort/size estimation is calculated using the proposed model online shopping system as a case study. The results indicate that the proposed model can help to estimate project size earlier in the design phase, to predict effort needed to complete development. The percentage of estimated effort between two diagrams is 85.49 is obtained .**

***Keywords*:** *Effort estimation, Use case points, Class points , Project size, Project complexity, Metrics.*

I. INTRODUCTION

Software effort is used to measure the use of the workforce. It is the total time that the members of a development team required to perform a given task . It is usually expressed in units such as man-hours , man-day , man-month , man-year. This serves as an indicator for estimating other values relevant , like cost [8].

An accurate estimation of effort is the most important factor in industries [7]. Both under estimation and over estimation can cause severe problems such the underestimation leads to under staffing and consequentially takes longer to deliver project than necessary. Over estimation may lead to miss opportunities to offer funds for other projects in future [9]. To avoid this, human experiments are needed to judge on the results, so the researchers try to develop models for accurate software effort estimation [7].

Line of code is a very important unit for time and effort estimations and many researchers denoted that a count in LOC depends on the degree of code reusing and can be accurate five times higher than anther estimate[5]. So, empirical studies have had an important role in the evaluation of tools, methods before they introduced in real software[5]. The UCP equation is composed of three variables: Unadjusted use case points (UUCP), The technical complexity factor (TCF), The environment complexity factor (ECF). The UCP method is versatile and extensible to a variety of development and testing projects. It is easy to learn and quick to apply [15].

In this paper the impact of UML diagrams on effort estimation is explored. The focus will be on the use case and class diagrams which are pure measures of size and they can establish an average implementation time of project development. It aims to:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- *Propose a method for estimating OO software size, complexity and effort needed for the software to be developed.*
- *Apply the above metrics for a project of on-line shopping, as a case study.*
- *Find the percentage of estimated effort between the two diagrams.*

This paper is organized into 6 Sections. The related work is described in Section 2. The effort prediction design, properties, metrics and tool architecture are described in Section 3. The case studies and results are described in Section 4. Finally, conclusion and future work are described in Section 5 and Section 6 respectively.

II. Related Work

Anda et al [1] estimated the software effort based on use case components and compute the total time in hours. Lavazza and Robiolo [2] showed the measurement-oriented UML modeling can support the computing effort based on functional size and complexity as independent variables. Sridhar [3] proposed knowledge based effort estimation for multimedia projects and concluded that the accuracy of effort estimation can be improved using knowledge rules. Harizi [6] defined parameters of class diagram with their importance, complexity and studied their impact on software size estimation.

Azzeh and Nassif [9] aim to study the potential of using Fuzzy Model Tree to derive effort estimates based on UCP size measure. Bardsiri and Hashemi [10] produced a brief review of well-known approaches from software effort estimation , classified as algorithmic and non-algorithmic techniques, summarized several models with some aspects impacting effort and concluded that each model has its own environment to be effective. Alves et al [5] described a case study based on function points with two teams that developed a software for a real customer to estimate the size and complexity of a software .Saroha and Saha [11] tried to answer questions dealing with factors that impacting effort estimation and to give guidelines for getting accuracy of estimated effort.

Kirmani and Wahid [12] studied 14 projects and applied proposed model in case of use case point and approved its improvements on estimated effort. Kirmani and Wahid also [13] observed scalability in technical complexity factor, project methodology in environmental complexity factor and their impact on estimated effort. Whigham et al [4] proposed a transformed linear model as a suitable baseline model for comparison of software effort estimation methods.

III. The Proposed Model

First, we required to draw the use case and class diagrams for a specific system in enterprise architect tool (Sparx Systems Enterprise Architect, a UML 2.1 based modeling tool for designing and constructing software systems) [16], then generate an XMI file for each diagram and use them as inputs to the effort prediction tool (EPT). Second, will be used a number of special metrics in software engineering to estimating size, complexity and effort of the project through the following steps:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Use case points

1) Computing use case complexity.

1.1) Calculate Unadjusted actor weight (UAW) by summation of a number of actors (NOA) multiplied with their weight.

1.2) Calculate Unadjusted use case weight (UUSW) by summation of a number of use cases (NOUC) multiplied with their weight

1.3) Calculate the number of roles (NOR)

2) Calculate unadjusted use case point (UUCP)

3) Calculate technical complexity factors (TCF) and environmental factors (EF) : We used Seventeen standard technical factors to estimate the impact on productivity and eight factors to estimate the impact on environment [14]. Each factor is weighted according to its perceived impact.

4) Calculate use case points (UCP)

*5) Assume that PF=20, then calculate Effort = UCP * PF .*

B. Class points

1) Computing class complexity.

1.1) Calculate the number of state point (SP) of class through summation the total functions in each Class multiplied by its own weight.

1.2) Calculate the behavioral point (BP) of class through summation the total number of Method in each Class multiplied by its own complexity and the result multiplied by one plus the number of associations per the class.

2) Calculate the number of class points in the project (CP) Size of class.

3) Calculate the size of each class in the class diagram .

4) Calculate the Size of a system .

5) Calculate effort based on the size of the system.

Figure 1 illustrates the steps of execution represented by an activity model for use case points and figure 2 illustrates the steps of execution represented by an activity model for class points. The enterprise architect tool is used to draw the models [16].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

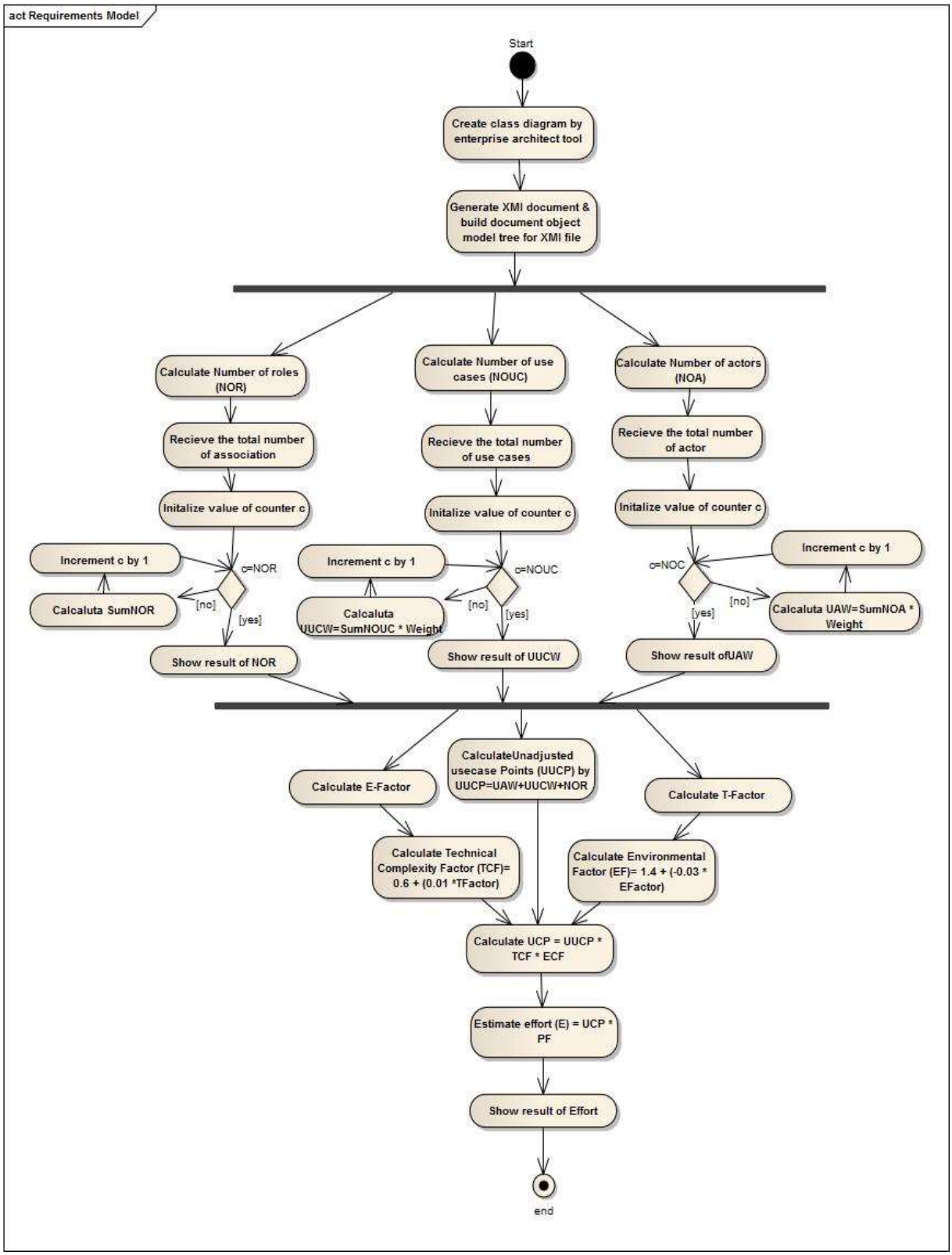


Figure 1. Activity model of use case pionts.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

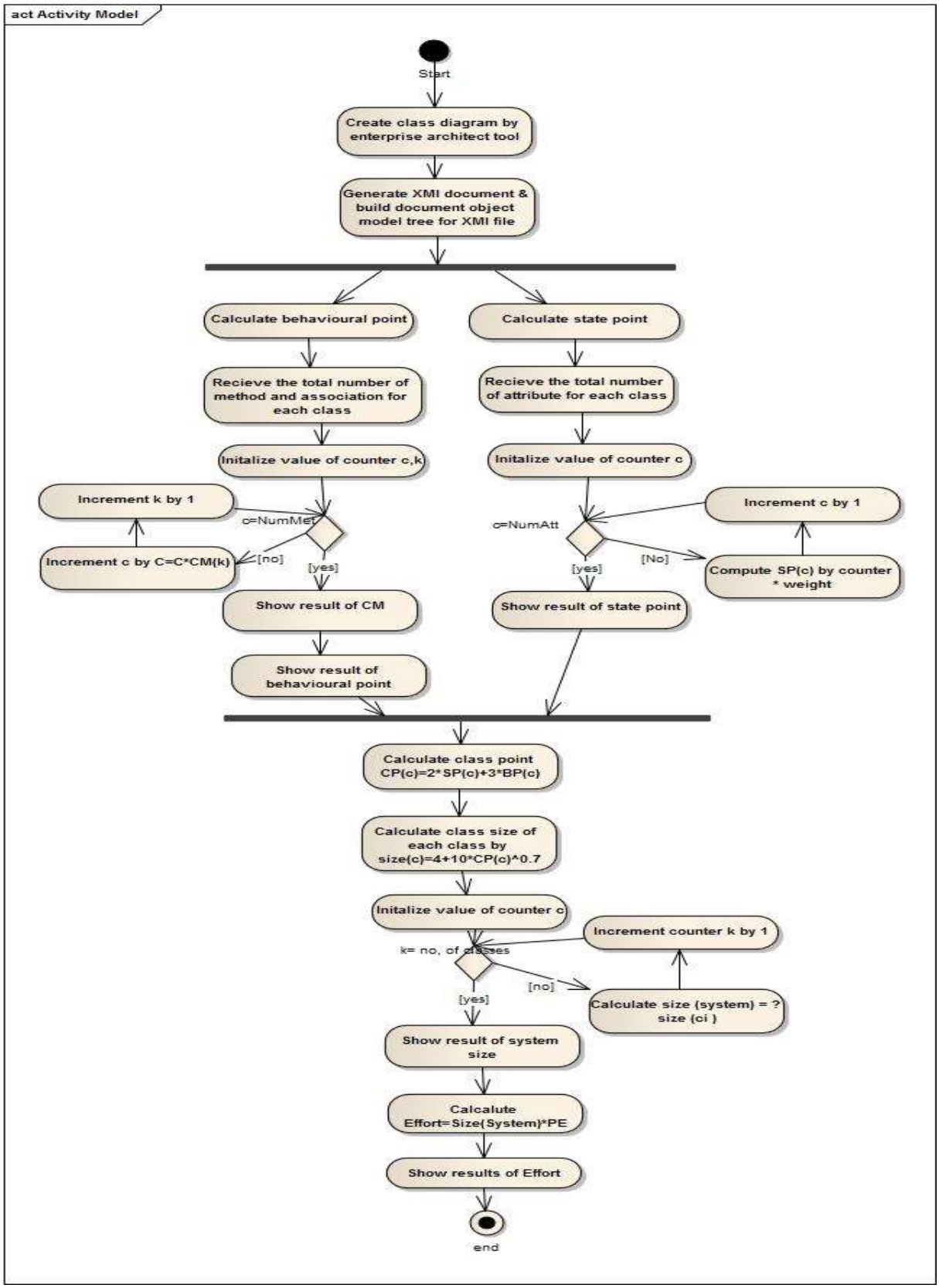


Figure 2. Activity model of class pionts.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. Proposed model Testing and Results

This section explains testing of the proposed model and results that are obtained after executed it.

4.1) Case Study

A case study of an online shopping system is taken and the metrics of the proposed model to obtain the size of project and effort needed. Figure 3,4 illustrate the use case and class diagrams for online shopping system and table1,2 illustrate the detail results obtained respectively.

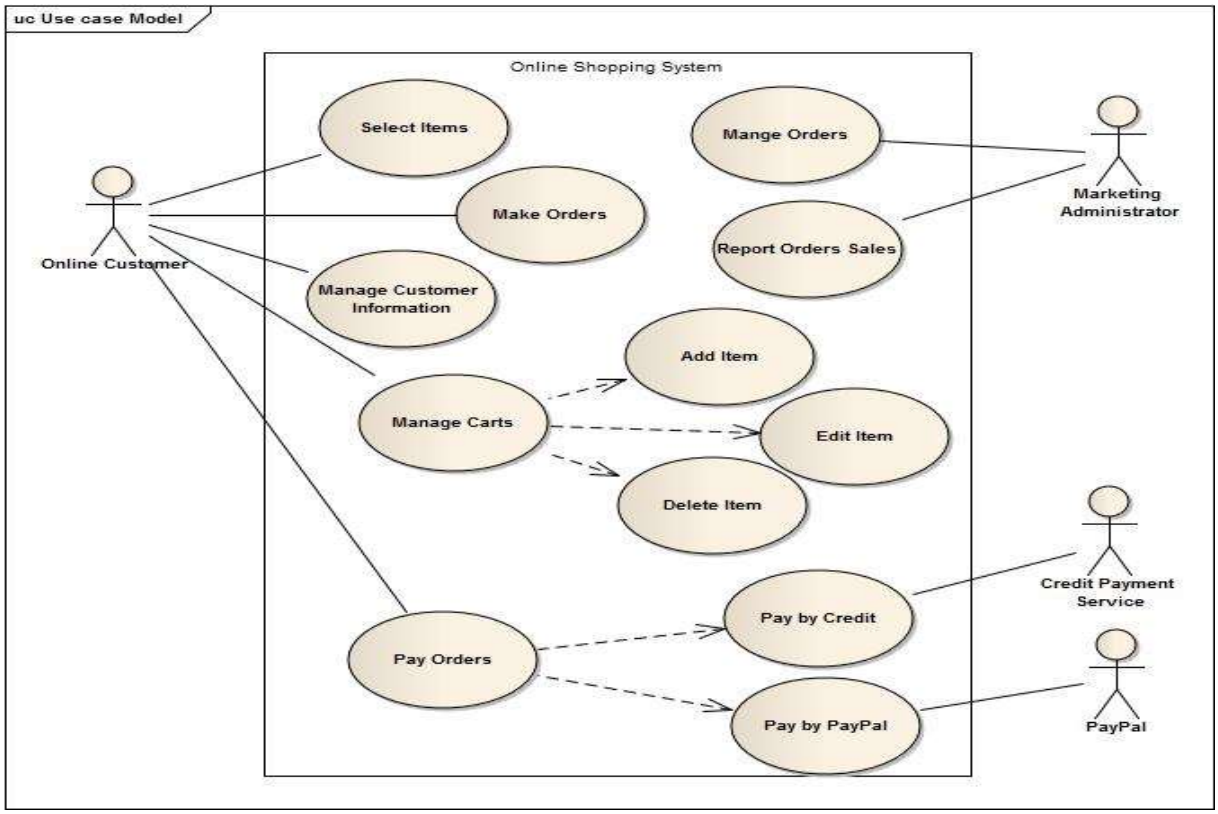


Figure 3. Use case diagram for online shopping system.

TABLE I
Results of use case diagram to predict effort

Description	Variables	Value
Number of actors	numActor	4
Number of use case	numUsecase	7
Number of rules	numRole	14
Unadjusted actor weight	UAW	10
Unadjusted role weight	URW	28
Unadjusted use case point	UUCP	98
Technical Complexity Factor	TCF	1.1
Environmental Complexity Factor	ECF	0.86
Productivity Factor	PF	20
Use case point	UCP	92.7
Effort	E	1854 man/hours

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

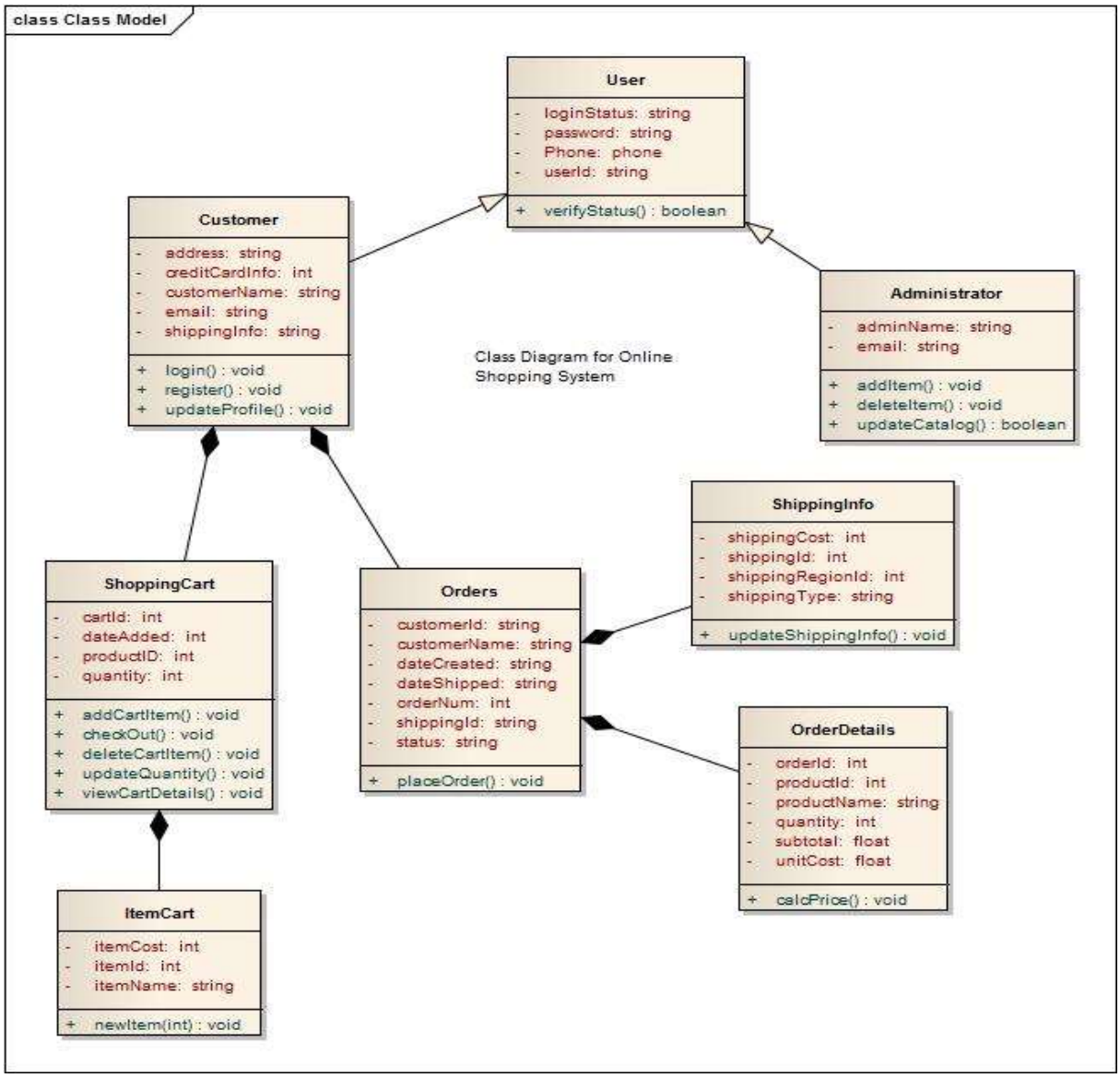


Figure 4. Class diagram for online shopping system.

TABLE II
RESULTS OF CLASS DIAGRAM TO PREDICT EFFORT

Class Name	SP	BP	CP	Size(Class)
User	4	3	17	76.66
Customer	5	12	46	149.86
Administrator	2	6	22	91.04
ShoppingCart	4	15	53	165.06
Orders	7	4	26	101.83
ShippingInfo	4	2	14	67.43
OrderDetails	6	2	18	79.63
ItemCart	3	2	12	60.94
Size(System)=792.45				
Effort=1585 man/hours				

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. CONCLUSION

This research presented a method for estimating OO software project size and the effort needed exploiting UML diagrams. So, through the building and testing of the proposed model, conclusions are:

- *The proposed model metrics can help software engineers to estimate project size and complexity in terms of lines of code earlier in the design phase.*
- *The proposed model can help to predict effort needed to complete development of the project easily in terms of man/hours and to give indicator for managing the overall budgeting and planning.*
- *The percentage of estimated effort between class and use case diagrams is 85.49 .*

VI. FUTURE WORK

In the future, the work may be enhanced in the following aspects.

- *Estimate effort of proposed model can be expanded using information extracted from sequence diagrams, activity and state chart or other diagrams of UML.*
- *The UML points can apply to more projects to provide guidelines for how to measure effort in different kinds of projects.*
- *The proposed model accepts only XMI documents generated by EA, so a model can be extended to accept XML documents also.*

REFERENCES

[1] B.C.D Anda, H.C Benestad, and S.E Hove,"A Multiple-Case Study of Effort Estimation based on Use Case Points", International Symposium on Empirical Software Engineering, ISESE'2005. IEEE Computer Society, Noosa Heads, Qld., Australia, Nov. 17-18, pp. 407–416, 2005.

[2] L. Lavazza,G. Robiolo, "Introducing the evaluation of complexity in functional size measurement: a UML-based approach", in: Proceedings of the 4th International Symposium on Empirical Software Engineering and Measurement – ESEM 2010, Bozen, Italy, 2010.

[3] S. Sridhar, "Extended Angel: Knowledge-based Approach for LOC and effort Estimation for Multimedia Projects in Medical Domain", International Journal of Software Engineering & Applications (IJSEA), Vol.2, No.4, October 2011.

[4] P. A. Whigham, C. A. Owen, andS. G. Macdonell, "A baseline model for software effort estimation". Journal ACM Transactions on Software Engineering and Methodology , ACM TOSEM,Volume 24 Issue 3, May 2015 PP 20:1–20:11, 2015.

[5] L. M. Alves, S. Oliveira , P. Ribeiro,R. J. Machado, "An empirical study on the estimation of size and complexity of software applications with function point analysis", 14th International Conference on Computational Science and its Applications; Guimaraes, IEEE 2014. p. 27–34.

[6] M. Harizi, "The Role of Class Diagram in Estimating Software Size", International Journal of Computer Applications, Volume 44, No. 5, pp. 31-33, 2012.

[7] K. U Sharani, V. Vignaraj Ananth., D. Velmurugan, "A survey on software effort estimation", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE, India, 2016.

[8] J. Zivadinovic, Z. Medic, D. Maksimovic, A. Damnjanovic, S. Vujcic, "Methods of Effort Estimation in Software Engineering", International Symposium Engineering Management and Competitiveness 2011 (EMC2011) June 24- 25, 2011, Zrenjanin, Serbia.

[9] M. Azzeh, and A. B. Nassif, "Fuzzy Model Tree For Early Effort Estimation", 12th International Conference on Machine Learning and Applications, DOI 10.1109/ICMLA.2013.115, IEEE computer society, 2013.

[10]A. K. Bardsiri, and S.M. Hashemi, "Software Effort Estimation: A Survey of Well-known Approaches", International Journal of Computer Science Engineering (IJCSE), 2014. 3(1): pp. 46-50.

[11]M. Saroha, S. Saha; "Analysis of various Software Effort Estimation Techniques", International Research Journal of Computers and Electronics Engineering (IRJCEE) Vol. 3, Iss. 2 , 2015.

[12]M. M. Kirmani, and A. Wahid," Revised Use Case Point (Re-UCP) Model for Software Effort Estimation", International Journal of Advanced Computer Science and Applications, 6, 65-71,2015.

[13]M. M. Kirmani, and A. Wahid, "Impact of Modifications made in Re-UCP on software effort estimation" Journal of software engineering and applications, Vol. 8, No. 6, pp276-289,2015.

[14] P. K. Bhatia, G. Kumar, " Role of Technical Complexity Factors in Test Effort Estimation Using Use Case Points ", International Journal of Software Engineering Research & Practices Vol.1, Issue 3, July, 2011.

[15] R. K. Clemmons, "Project Estimation With Use Case Points", The Journal of Defense Software Engineering, February 2006.

[16] http://sparxsystems.com.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Design and implementation elliptic curve digital signature algorithm using multi agent system

Tawfeeq M. Tawfeeq Al-Flaih
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
Tawfeek.mflaih@yahoo.com

Marwa Adeeb Al-jawaherry
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
marwaaljoharjy123@gmail.com

Abstract— **The rapid and growing development of information and communication technologies ICTs, especially in the Internet, has been a key driver for improving the quality and efficiency of services provided by many countries. The digital signature algorithm (DSA) is designed to dispense with the signature in handwriting and replace it with a signature, and it helps us to verify the identity of the sender and receiver in a reliable and secure manner.**

In this research we are proposing and constructing security system which is called Digital Signature Multi Agents (DSMA). It is based on Multi Agent System (MAS) and provides authentication of senders or receivers by applying "Elliptic Curve Digital Signature Algorithm (ECDSA)" to sign and to verify the electronic documents. Two types of agents were developed in our proposed system: sender and receiver agent. Java programming language and JADE (Java Agent Development framework) were used to constructing DSMA.

Keywords- Hashing Algorithm; Elliptic Curve Cryptosystems; Elliptic Curve Digital Signature Algorithm; Multi Agent System; Java Agent Development

I. INTRODUCTION

With the increasing of the online application and electronic transactions, The transition from paper based transactions to electronic transaction become more easy and less complicated but the challenge lies in the implementation of these transactions in terms of the validation and insurance. It can be viewed to the digital signature technology as a mechanism to maintain the integrity and safety ratio in electronic transaction [10]. As the dependence on the Internet for the exchange of information and communication continues to increasing, the security concerns are becoming more important. There is desperately need for digital identity or digital signature, which will activate the quality of our dealings and contacts increase the security [17].

It is noted that when conducting transactions electronically, there is no way to confirm the identity of the sent or received transections, hence the possibility of the use of digital signatures to authenticate the source of the electronic messages or transactions; The digital signature confirms the true identity of the sender, and more importantly, can be used to maintain the integrity data, from editing, sources posing a strength and excellence, for these reason digital signatures is an effective solution for authentication and documentation [14][23].

In this paper we built and developed a secure system based MAS called Digital Signature Multi Agents (DSMA) that sends text messages to many distributed sites, and implement a Digital signature algorithm (DSA) to verify the integrity of sent and signed data, in addition to verifying sender identity.

DSMA can be executed on any system that rely style electronic exchange of official documents, It is possible to apply DSMA system for the exchange of official document traded electronically between the presidency of the University of Mosul and between different colleges or for the exchange of official document electronically between Deanship colleges and departments. DSMA system can accept any number of users, the user can be divided into two types:

A. Sender

The person who is sends electronic documents after the process of generating a digital signature and encryption of the document and then sends it to the receiver site by his personal agent.

B. Receiver

The person who is receives electronic documents that have been digitally signed and transmitted by sender agent.

It is worth mentioning that all users of DSMA system can be mailed electronic documents at the same time.

II. MOTIVATION

The digital signature (DS) is a mathematical method to clarify whether the digital messages or documents received are true or not. It also helps the receiver to verify the validity of the sender's identity (authentication), in this case, the sender cannot deny sending the message (non-repudiation), and can be sure that the message is not changed during the transfer process (integrity).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

DS is one of the standard elements in most suites of the cryptographic protocol, DS is used in many areas, such as financial dealing, software used in distributions and contract management systems, As well as systems to detect tampering or counterfeiting. Multi-agent systems are used in many areas such as network security systems [5].

Our research aims to develop security system based on MAS, We tried to overcome some of the difficulties we faced in this paper, recognize the authentication of the signatory and using multi agent system to implement. Our system has several requirements that have to be achieved in his work as follows:

1. The main function of the system is to make sure of the identity of the sender of confidential information after receiving it.
2. The number of system user is different and is not specific to a certain number, for example there may be four or five users.... etc., and so the number of computers linked to the network is not specified.
3. Any user of the system can send information over the network to the rest of other users through the software agent.
4. Every user of the system has a personal agent represents him and who interacts with each others by sending messages.
5. Any user can be the sender and receiver of the messages in the future and at the same time.

III. RELATED WORK

As long as people have been able to communicate with one another, there has been a desire to do so secretly. Many researchers work with digital signature algorithm, Cloud Computing and MAS:

1. In (2011), Erfaneh Noorouzi1 and his colleagues proposed a new DSA algorithm, which generates dynamic size hash files, which mean the size of the message affects the result of the hash function. The mechanism for (hash / encrypt) will be more simple by a new DSA algorithm [18].

2. In (2011), Aarti Singh and his colleagues proposed "security engine" to secure messages sent in networks environment, this proposed make Elliptical Curve keys used for the purpose encrypt and decrypt. This framework can be implemented in the security layer of the current wireless communication model for this reason is not needed to rewrite it to use [29].

3. In (2012), Salwanibtmohd Mohd daud and his colleagues produced DS from achieved a simple mechanism by proposing a new algorithm. The resulting output would be dynamic and smaller by this new algorithm. Hashing and encoding the message after the algorithm read the input file [19].

4. In (2012), Thulasimani Lakshmanan and Madheswaran Muthusamy used SH Algorithm to present a new SHA called "SHA-192". The output length message "SHA-192" of 192. They designed "SHA-192" to resist the SH Algorithm attacks and to fulfill the different level of information security[14].

5. In (2016), Virangna Pal and his colleagues discussed the two types of Security algorithms (Symmetric and Asymmetric Algorithms) that were used in "Cloud Computing", they checkup various constraints for ex: features and mechanisms, and they discussed some case connected with distributed system [30].

IV. METHODS & MATERIALS

A. *Electronic Signature*

An electronic signature refers to data that has an electronic format, which is associated to other electronic data logically, and this data will be used by the signatory to perform signing process. The main objective of applying the electronic signature process is to provide accurate and safe way to verify the identity of the sender. It is worth noting that the definition of electronic signature depends on the jurisdiction that Applied. There are three types of electronic signature, as following: Digital Signature, Personal Signature and Signature Using Pen Mail [4] [16].

Digital signature is an encryption process is composed of some of the letters, symbols and numbers. It can be represented as a string of binary digits in a computer, and must achieve the functions where the signature identifies the signer's identity and the expression of his will approve the content of the message data [23][12]. The digital signature value is calculated using a number of parameters that verify the integrity of the signed data and the identity of the signatory [10] [13]. The digital signature having several requirements includes UN forgeable, User authentication, Non-repudiation, Unalterable and Not reusable [20].

B. *Software Agents*

Agents are separate pieces of software that have the ability to act independently and interact with the environment in which they operate. There are different types of agents so their abilities are also different. In order for the agent to be described as an "intelligent" agent, he should have the ability to interact with other agents or with his environment without the need for direct interaction by human beings as well as must be flexible [6].There are four types of agents: Executive agents, Collaborative agents and Contributory agents [9][25].

Multi-agency systems are modern approaches to analysis, design and implementation of complex software systems. To develop and implement different types of software systems, it is possible to use multi-agent systems and is also used in the development of search and rescue systems and network security [11]. MAS are used to describe several agents that interact with each other positively, but also negatively within an environment [22][8].

C. *JADE platform*

JADE platform is a software framework port language Java. It was developed by the Research Institute of the Italian contact in 1998 by using a set of graphical tools to simplify the implementation of multi-agent systems [26]. The goal of JADE is to facilitate the development and to ensure that the standard

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

response by providing a set of services for the overall system, as well as providing a variety of agents [27].

JADE architecture consists of agent containers that are on the same platform but distributed over the network. Each agent lives in a container which is a Java process that provides a JADE runtime and all services necessary to host and execute agents. In each platform there is a special container, called the main container, which is launched at the platform and which contains the other containers in which they are registered [22]. The interaction is the most important properties of the agent, and the agent interaction to share information and knowledge in order to achieve his goals. In order every agent to own a mechanism to achieve compatibility, there are two key elements in the agent connections: Protocol negotiations common/ language of communication and Representation of the general formula content [27].

D. Secure Hash Algorithm (SHA)

Secure Hash Algorithms provides many services while used in other cryptographic algorithms [1]. Converting a variable length message into a condensed representation of the electronic data in the message is made by Hash algorithm, and this output can then be used for DS and any other secure system. When employed this representation in a DS application, the "Hash value" of the message is signed instead of the message itself, then the receiver can authenticate the integrity of the signed by using the signature to verify the signer of the message [3][18][21].

E. Digital Signature Algorithm (DSA)

The U.S.NIST in August 1991 proposed "Digital Signature Standard (DSS)" [23]. The key generation process consists of two stages, choosing algorithm parameters that can be shared between the various users of the system, in the first stage. The second stage involves calculating the private and public keys for the same user [17].

F. Elliptic Curves Cryptosystems (ECC)

In 1985 Neal Koblitz and Victor Miller invented ECC. It can appear as EC analogues of the older Discrete Logarithm crypto systems [15]. Public key cryptography based on the algebraic structure of EC over finite fields. ECC requires smaller keys compared to any other cryptography to provide equivalent security. ECC are applicable for (key agreement, DS, and other tasks), they can be used for encryption by combining the key agreement with a symmetric encryption scheme and used in several integer factorization algorithms based on ECC [28].

G. Elliptic Curve Digital Signature Algorithm (ECDSA)

The ECDS Algorithm is the "Elliptic Curve" analogue of the commonly used DSA [20]. ECDS Algorithm offers technical avail in the areas of certificate, performance, and key over other DS methods [24], Figure (1) shows the interaction between ECDS Algorithm and SHA-2 [7]. RSA or DSS are very difficult or expensive to implement in specific applications while smaller data structures and calculation efficiencies for ECDS Algorithm enable it to be used in these applications [13][16].

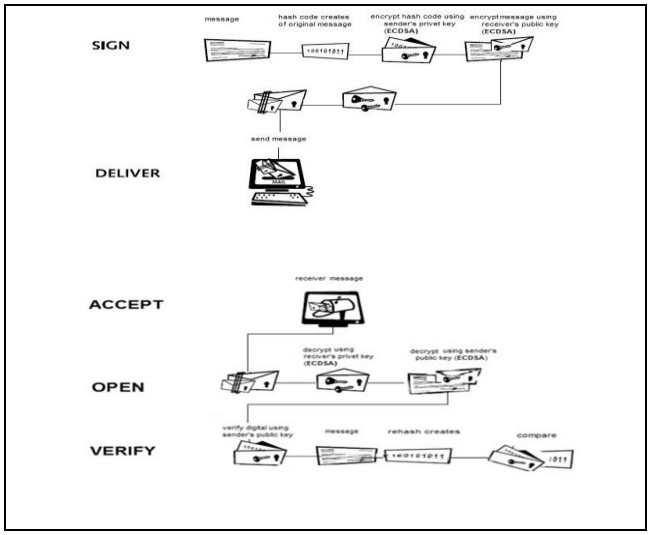


Figure 1. Represent interaction between ECDSA and SHA-2

V. THE LIFE CYCLE OF PROPOSED WORK

This paragraph presents clarification of the proposed DSMA system, which is designed to ensure and confirm the identity of the sender of electronic documents, as well as making sure it from the correct source. We will describe the system architecture, as well as explain JADE interfaces that are used to communicate with the system and clarify all code components. Also we will indicate the number of agents in DSMA system and user characteristics and responsibilities. We will use the Smart MAS style to analyze, the design and implementation of our system.

A. Requirements phase

After the initial analysis of the requirements, the representation of the active ingredients in the simple scheme actor. In our proposed system we have two types of actors, first: sender of the messages, second: receiver of the messages. The main objective of the sender is to generate digital signature and send messages. While the main objective of the receiver is receiving the messages and confirming the identity of the sender, Figure (2) shows the actors of a simple scheme for DSMA system.

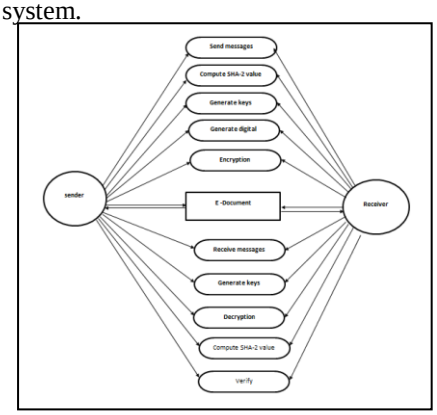


Figure 2. Simple scheme actor of the DSMA system

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Advanced requirements consist of four steps: insert system actors, creating goals diagrams, creating actor diagrams and analyze dependencies.

1. Insert system actor: in this step the system actors are inserted under development in a simple diagram and its own tasks are appointed as shown in Figure (3), which shows the system actors that have been delegated all the goals except the resources which are outside the system.

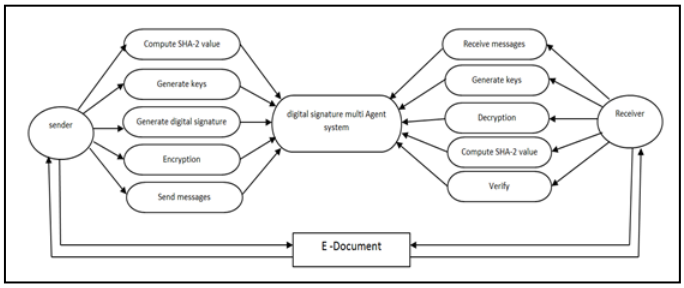


Figure 3. illustrates the DSMA system Actor

2. Creating goals diagrams: This step is centered in three sub-phases, As explained In Figure (4) and Figure (5). show the goals decomposition of the sender and receiver.

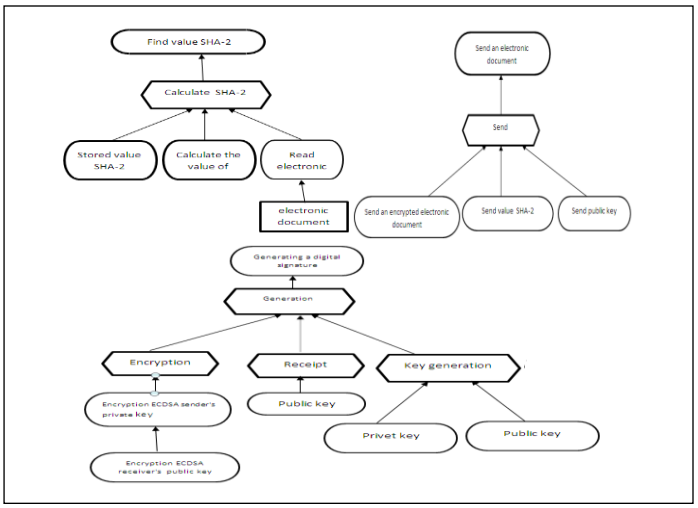


Figure 4. Goals decomposition of the sender

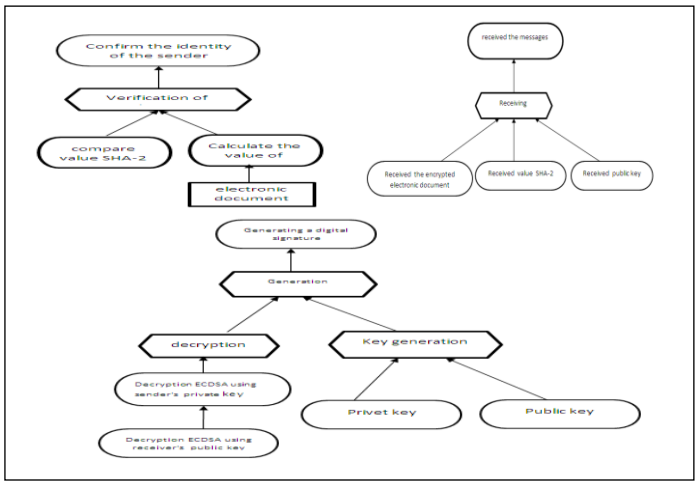


Figure 5. Goals decomposition of receiver

3. Creating actor diagram: After assembling the plans for system actors and goals, final actor diagram is formed for requirements phase, Figure (6) shows the final actor diagram.

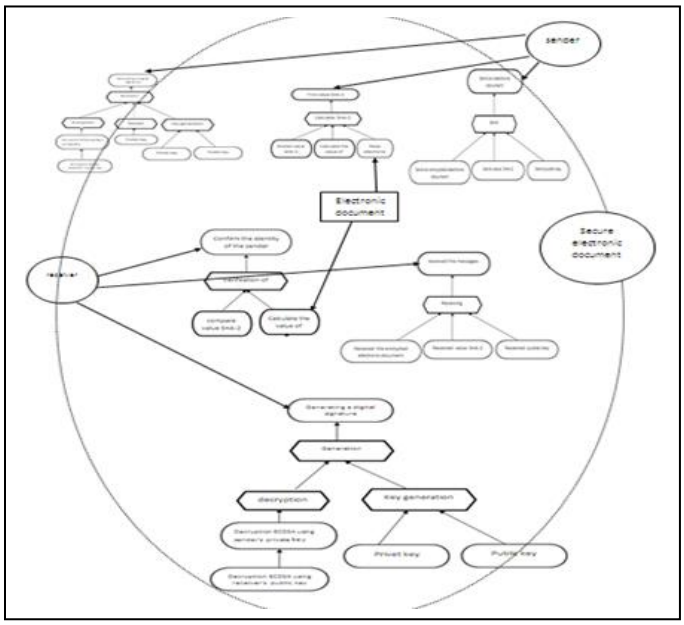


Figure 6. Final actor diagram

4. Analyzed dependencies: it was analyzed between actors who are (sender, receiver) and DSMA system as shown in table (1).

B. *Analysis Phase*

The analysis phase is divided into two major steps: the first is to create a structure description of the dependencies between agents, see table (1). And the other step is a description of the role for each agent in DSMA system, see table (2).

TABLE I. ANALYZED DEPENDENCIES BETWEEN ACTORS IN DSMA

Sender dependencies It is needed sender actor to achieve its goals, as in the following formulas: **Dependency**: find the value of hash **Dependent**: sender. **Dependee**: receiver. **Dependum**: Electronic document. **Goal**: calculate the hash value by implementing the algorithm sha-2 (384-bit). **Pre-condition**: the presence of the electronic document. **Post-condition**: determining the validity of information and use it to generate DS. **Dependency**: the generation of the DS. **Dependent**: sender. **Dependee**: receiver. **Dependum**: Electronic document. **Goal**: generates a digital signature for electronic document. **Pre-condition**: the presence of the electronic document and find the value computed hash. **Post-condition**: generating each of the (public, private) keys and receiver of receiver public key, and encrypt electronic document **Dependency**: electronic document using ECDSA. **Dependent**: sender. **Dependee**: receiver. **Dependum**: Electronic document. **Goal**: encrypt electronic document. **Pre-condition**: the existence of a signed electronic document. **Post-condition**: Send electronic document to the receiver. **Dependency**: Sends an electronic document. **Dependent**: sender. **Dependee by**: receiver. **Dependum**: Electronic document. **Goal**: electronic document delivery to the receiver. **Pre-condition**: the existence of an encrypted electronic document and provide contact with the receiver. **Post-condition**: Posted / / Not transmitter.	**Receiver dependencies**: It is needed receiver actor to achieve its goals, as in the following formulas: **Dependency** : receive of messages. **Dependent**: receiver. **Dependee**: sender. **Dependum**:messages. **Goal**: receive of the electronic document and the public key and the value of the SHA-2. **Pre-condition**: the presence of the electronic document. **Post-condition**: has been receiving / / No receive. **Dependency**: generate keys **Dependent**: receiver. **Dependee**: sender. **Dependum**: generate the (public, private) key . **Goal**: electronic document will be encrypted using the public key by the sender and use the private key to decrypt the electronic document by the receiver. **Pre-condition**: the presence of the receiver. **Post-condition**: Send the public key to the sender. **Dependency**: decryption of electronic document . **Dependent**: receiver. **Dependee**: sender. **Dependum**: message. **Goal**: decrypt message using ECDSA. **Pre-condition**: receive the electronic document from the sender **Post-condition** : perform verification algorithm. **Dependency** : calculate the value sha-2 (384-bit). **Dependent**: sender. **Dependee**: the receiver. **Dependum**: Electronic document. **Goal**: find the value of SHA-2. **Pre-condition**: electronic document encryption using ECDSA. **Post-condition**: compare the value of sha-2 calculated with the value of SHA-2 received from the sender. **Dependency** : confirm the identity of the sender of the electronic document . **Dependent**: receiver. **Dependee**: sender. **Dependum**: electronic document. **Goal**: confirm the identity of the sender by verifying the digital signature . **Pre-condition**: compare the value of sha-2 with the calculated value of the SHA-2 received from the sender. **Post-condition**: Sender is trusted / / sender is not trusted.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II. CLARIFIES THE ROLE OF THE SENDER AND RECEIVER

The role of the sender	The role of the receiver
Description: This role is a process of sending electronic document after calculating the SHA -2 , encrypted it using ECDSA algorithm and confirms the identity of the sender when it is sents . **Main Goal**: generates a digital signature and send electronic document . **Dependency** : Send an electronic document . **Activities**: receives the receiver's public key and send electronic document , application two algorithms SHA-2 , and ECDSA and configure agents. **Successful actions**: the generation of the digital signature . **Failed actions**: the inability to generate a digital signature, send electronic document , and verify the identity of the sender.	**Description**: This role is a process of receiving electronic document which is sent by the sender, decryption using ECDSA algorithm, and calculate the value of the SHA-2 to make sure of the identity of the sender of the electronic document . **Main Goal** : receives the electronic document , and confirm the identity of the sender. **Dependency** : receive the electronic document , the sender's public key , the value of the SHA-2. **Activities** : receives electronic document , send the public key of the receiver to the sender ,applying SHA-2, and ECDSA algorithms, comparison and configure agents. **Successful actions:** the exchange of messages and communicate with the sender. **Failed actions**: the inability receive electronic document . and can't verify the

There are two types of agents, each one with a specific role in the DSMA system Figure (7) illustrates agents and its own role.

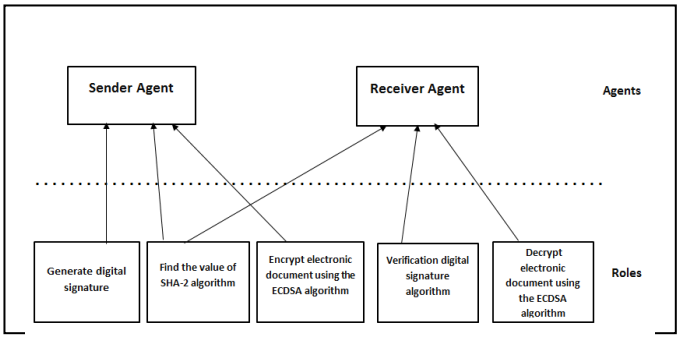


Figure 7. Agents and its own role

C. *Design Phase*

After the defining of agents and setting goals and their own tasks, we can create a plan to deploy these agents in locations that can be found there in, as well as a description of its functions. The proposed system contains two of the agents who are the sender and the receiver and the note through the requirements, it should be distributed and can be There a different numbers of senders and receivers, On this basis deployment scheme has been configured the presence of the sender in the platform and the agent of the receiver in another platform exist on the same network and the number of copies of the sender agent and receiver agent is not specified because it is dependent on the number of users of the system. As a software engineer we are focus on implementing the most important design concepts that allied "Modularity" which divided software to components, each component has its own name and address that called "Modules". Figure (8) show the control hierarchy and the modules in DSMA system.

D. *Construction Phase*

DSMA system consists of one package that contains the following classes: Sender Agent Class, Receiver Agent Class, Encryption Class, Decryption Class, Digital Signature Class, Hashing Class. We built the DSMA system using the JADE framework under Java language.

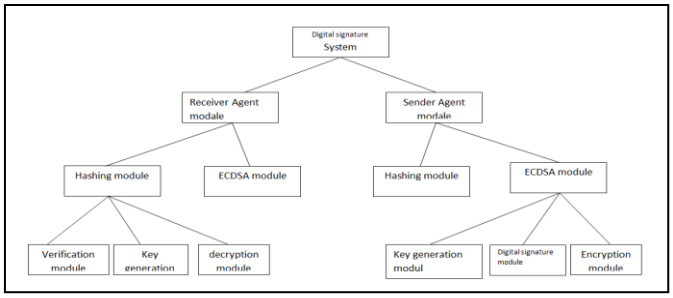


Figure 8. control hierarchy and modules in DSMA system

VI. CASE STUDY

This section presents the using and testing of DSMA system. It implemented practically and discussed according to the results obtained.

The proposed DSMA system is a distributed system, so the implementation needs to provide a number of computers linked with each other through the LAN, and the number of those computers is not specified, and each user can interact with other users through the agent who represented him and the agent will keep working in his computer.

In the beginning, DSMA system is used by running the JADE platform at the sender and receiver sides, then the user should create sender and receiver agent on each platform. The sender agent reads the electronic document and finds the value of the hash code using algorithm SHA-2 (384-bit) and sends it to receiver, the keys and digital signature are to be generated. Then encrypts electronic message and sends it also to receiver. The receiver agent receives electronic messages and decrypts it, finds the value of hash code using the algorithm SHA-2, compares the value of hash calculated with the value of the hash received, if the value is equal, it means that the document is received from the correct sender, but if it is not equal, it means that the document is received from the incorrect sender. On this basis, assurance of the identity of the sender is achieved. Figure (9) and figure (10) show the JADE interface at Sender / Receiver site.

Agents are interacting with each other by exchanging ACL messages, several types of behaviors are used in the implementation of the tasks of the agent, namely: One shot behaviors and cyclic behaviors. Sender and receiver agents can be resident in the main or secondary containers in JADE platform.

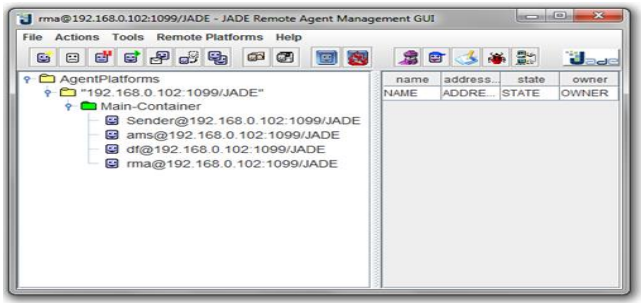


Figure 9. Sender agent.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

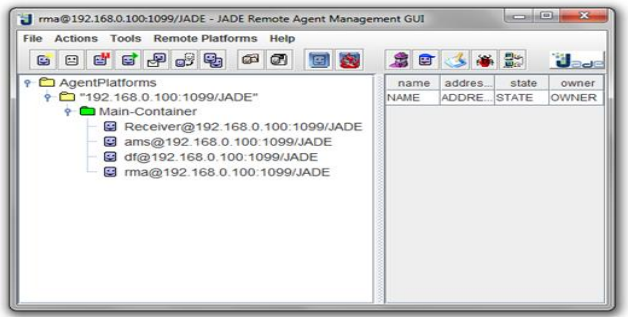


Figure 10. Receiver agent

DSMA system includes two types of agents:

A. The sender's agent:

1. (Key Generation Algorithm):

At this step the public key and private key are generated. Later keys are to be used in encryption and decryption operations.

```
.
.
public class Key_Generate {
        public static String[] point(String s1, String s2 , String l)
        {  BigInteger a,b,d,x1,y1,s,aa,bb,v1,v2,c,c1,x,y;
                                .
                                .
                                .
public static String[] key_generater_pri(String private_key) throws
UnsupportedEncodingException
        {
          BigInteger q1,q2;
          String[] result_A;
             result_A = point("0","2",private_key);
             q1= new BigInteger(result_A[0]);q2= new
BigInteger(result_A[1]);
              String ar[] = new String[2];
             ar[0]=q1.toString();
         ar[1]=q2.toString();
           return ar;
         }
}
```

2. (Signing Algorithm):

By using this algorithm the sender agent generate a digital signature for electronic document and implicitly calculates the value of the SHA-2.

3. (Key exchange operation):

The sender agent sends his public key to receiver agent.

```
public class Sender extends Agent
{ String str,hash,result[],result1[],result2[];
String result_pub_A[],result_pub_B[],priv_A,priv_B;
protected void setup()
                        {
                                System.out.println("Enter  String  :
");
                                Scanner       sc11       =       new
Scanner(System.in);
                            str = sc11.nextLine();
                        System.out.println("Enter Private Key A : ");
                        priv_A=sc11.nextLine();
                                try {
 result_pub_A=Key_Generate.key_generater_pri(priv_A);
                                }
                        System.out.println("public
key"+result_pub_A);
 addBehaviour(new CyclicBehaviour(this)
                        {               public void action() {
                                        ACLMessage msg1rec= receive();

                                                if (msg1rec!=null){
                                                        String title =
msg1rec.getContent();


        System.out.println( " - " +

myAgent.getLocalName() + " <- " +

msg1rec.getContent() +"    " + title);


                                                block();
                                                }
                                }
                };
```

4. (Encryption Algorithm):

At this step electronic document was encrypted by applying ECDSA algorithm before sending them to the site of the receiver, the encryption of electronic document contain two operations: First, encrypted electronic document by using private key of the sender , second, encrypt the electronic document by using the public key of the receiver.

```
public class encryption {
        public static String[] point(String s1, String s2 , String l)
        {
                        BigInteger a,b,d,x1,y1,s,aa,bb,v1,v2,c,c1,x,y;
                        BigInteger p = new BigInteger("137");
                        aa= new BigInteger("3");
          bb= new BigInteger("2");
          c= new BigInteger("4");
          c1= new BigInteger("27");
          a = new BigInteger("1");
          b = new BigInteger("4");
          x = new BigInteger(s1);
          y = new BigInteger(s2);
          d = new BigInteger(l);
          int ss = p.intValue();
          ss=ss-2;
                        x1=x;
          y1=y;
                        for(int i=2;i<=d.intValue();i++)
          {
                                if (x1==x&&y1==y)
            {

        v1=x.pow(2).multiply(aa).add(a);
                                v2=bb.multiply(y);
                                v1=v1.mod(p);

s=v2.pow(ss).multiply(v1);
                                        s=s.mod(p);

x1=s.pow(2).subtract(x1).subtract(x);
                                x1=x1.mod(p);

y1=s.multiply(x.subtract(x1)).subtract(y);
                                y1=y1.mod(p);
                }
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

```
                                        .
                                        .
                                        .
public static String[] encrypt(String str1 ,String x,String y,String pri)
throws UnsupportedEncodingException
        {
         String[] result;String[] result_key,decrypt,re = null;
        String cipher2 = null;
         String stre = null,strd = null,ad;ad=" ";
         int[]a;int c=0;int e=0;
         int l;
         l=str1.length();
         if(l%2==0)
         {
                        System.out.print(" ");
         }
         else
         {
                        str1=str1+ad;
         }
         BigInteger q1,q2,c1,c2,s1,s2,d1,d2,v,p,c21,c22,kk;
         v = new BigInteger("-1");
         p = new BigInteger("137");
         int co=0;int er=0;
         byte[] bytes = str1.getBytes("US-ASCII");
         l=str1.length();
         result_key=point(x,y,pri);
         s1=        new        BigInteger(result_key[0]);s2=        new
BigInteger(result_key[1]);
         BigInteger r,r1;String strre = null;
         String re1 = null;
         String[] re2;
         while(c<l)
         {              r = new BigInteger(Byte.toString(bytes[c]));
                        r1 = new BigInteger(Byte.toString(bytes[c+1]));
        cipher2=add2point(r.toString(),r1.toString(),s1.toString(),s2.to
String());
                        strre=strre+cipher2;
                        c=c+2;
         }
         strre=(String) strre.subSequence(4, strre.length());
         strre=(String) strre.subSequence(0, strre.length()-1);
         String[] bytes1 = strre.split(",");
         System.out.print("Encryption:");
         for (int a1=0;a1<bytes1.length;a1++)
         {        stre=stre+bytes1[a1];

System.out.print((char)Integer.parseInt(bytes1[a1]));
         }
          String ar[] = new String[4];
           ar[0]=stre;
           ar[1]=strre;
                        return ar;
         }}
```

TABLE III. ILLUSTRATE THE TYPE OF AGENT, MESSAGES INFORMATION (TYPE, NUMBER, AND CONTENT) IN SENDER AGENT.

Name of agent	Type of agent	No. of sending messages	No. of receiving messages
Sender	Static	3	1
In	**Messages Content**		**Message type**
1.	Public Key		Send
2.	Value of SHA-2		Send
3.	Signed electronic document		Send
4.	Receiver Public Key		Receive

B. The receiver's agent:

The receiver agent is responsible for receiving and decrypting electronic document and confirms the identity of the sender using ECDSA algorithm which is implicitly consists of three algorithms that are implemented at the receiver site as follow sequence, as can be seen in table (4).

1. (Key Generation Algorithm):

Public key and private key are generated. Keys will be used in encryption and decryption operations.

2. (Key exchange operation):

The receiver agent sends his public key to sender agent.

3. (Decryption Algorithms):

The decryption of electronic document contains two operations: First, decrypting electronic document by using private key of the receiver, second, decrypting the electronic document by using the public key of the sender.

4. (Signature verification algorithm):

Through using this algorithm we can ensure the authenticity of the digital signature after decrypted, and find the hash value of electronic message, then compare it with the hash value received from the sender's agent to ensure the identity of sender.

```
public class Receiver extends Agent
{
       String str,hash,result[],result1[],result2[];
       String result_pub_A[],result_pub_B[],priv_A,priv_B;
       private static final long serialVersionUID = 1L;

       protected void setup()
   {
                      Scanner sc22 = new Scanner(System.in);
                      System.out.println("Enter Private Key B : ");
    priv_B=sc22.nextLine();
    try {
           result_pub_B=Key_Generate.key_generater_pri(priv_B);
        }
    System.out.println(result_pub_B);

    addBehaviour(new OneShotBehaviour(this)
                      {
                       public void action() {
                       ACLMessage        msg2rec        =        new
ACLMessage(ACLMessage.INFORM);
                              final     String                    s1=
result_pub_B[0]+"**"+result_pub_B[1];
                                msg2rec.setContent(s1);


msg2rec.addReceiver( new AID( "s" , AID.ISLOCALNAME) );
                                send(msg2rec);
                                }
                      });
```


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE IV. ILLUSTRATE THE TYPE OF AGENT, MESSAGES INFORMATION (TYPE, NUMBER, AND CONTENT) IN RECEIVER AGENT.

Name of agent	Type of agent	No. of sending messages	No. of receiving messages
Receive	Static	1	3
In	**Messages Content**		**Message type**
1.	Public Key		Send
2.	Value of SHA-2		Receive
3.	Signed electronic document		Receive
4.	Receiver Public Key		Receive

VII. CONCLUSION

The present study, proposed a secure multi agent system (DSMA), and many points have been concluded besides the following:

1. The software agent has an ability to execute complex algorithms, and excellent results were got .

2. The ssoftware agent was a very good choice to execute numerical algorithms efficientlyy.

3. MAS has an ability to reduce the communication problems because the of low size of agent's messages.

4. The use multi agent system help to perform complex interaction between distributed sites.

VIII. ACKNOWLEDGMENT

We thank the in charge of laboratories of the Software Engineering Department - Mosul University for their help in testing DSMA system propose.

IX. REFERENCES

[1] Zhu, H., and Li, D.,(2008), "Research on Digital Signature in Electronic Commerce", Internationa MultiConference of Engineers and Computer Scientists, 19-21 March,Vol.1, IMECS 2008, Hong Kong.

[2] Bellifemine, F., Poggi, A., and Rimassa, G.,(2001), "Developing multi – agent systems with a FIPA-compliant agent framework", Softw:Pract.Exper., 31:103 –128.doi:10.1002/1097-024X(200102)31:2<103::AID-SPE358>3.0.CO;2 O.

[3] Bellovin, S., and Rescorla, E.,(2005), "Deploying a New Hash Algorithm", Technical Report, National Institute of Standards and Technology (NIST).

[4] Bettelli A.V.,(2004), "The Trustworthiness in Software Agents' Electronic Signatures".

[5] Borselius, N.,(2003), "Multi-agent system security for mobile communication", PhD Thesis submitted to the Department of Mathematics Royal Holloway, University of London.

[6] Borselius, N,.(2002), "Mobile Agent Security", Electronics & Communication Engineering Journal, Volume: 14, Issue: 5.

[7] Curry, I.,(2001), "An Introduction to Cryptography and Digital Signatures", Version 2.0, Entrust Securing Digital Identities & Information.

[8] Dale, J.,(1997), "A Mobile Agent Architecture for Distributed Information Management", PhD thesis, University of Southampton.

[9] Fabio, B., Caire, G., and Greenwood, D.,(2007), "Developing Multi-Agent Systems with JADE", John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, ISBN: 978-0-470-05747-6.

[10] Gallagher P. and Director D.,(2009), "Digital Signature Standard", National Institute of Standards and Technology, Gaithersburg, MD 20899-8900 , FIPS PUB 186-3.

[11] Hegde, M., and Singh S.,(2013), "Alert-BDI: BDI Model with Adaptive Alertness through Situational Awareness", IEEE. Conference Mysore, India. ISBN: 978-1-4673-6217-7

[12] Johnson D., Menezes, A., and Vanstone, S.,(2001), "The Elliptic Curve Digital Signature Algorithm (ECDSA)", International Journal of Information Security, Volume 1, Issue 1, pp 36–63

[13] Khalique, A., Singh, K., and Sood, S.,(2010), "Implementation of Elliptic Curve Digital Signature Algorithm", International Journal of Computer Applications, (0975 – 8887), Volume 2 – No.2.

[14] Lakshmanan, T., and Muthusamy, M.,(2009), "A Novel Secure Hash Algorithm for Public Key Digital Signature Schemes", The International Arab Journal of Information Technology, Vol. 9, No. 3, May 2012.

[15] Leslie, M.,(2006), "Elliptic Curve Cryptography", 40575063, Advanced Combinatorics.

[16] Liao, H., and Shen, Y.,(2006), "On the Elliptic Curve Digital Signature Algorithm", Tunghai Science, Vol. 8: 109−126.

[17] Muthukuru, J., and Sathyanarayana, B.,(2012), "A Secure Elliptic Curve Digital Signature Approach without Inversion", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-3, Issue-2.

[18] Noorouzi1, E., HAGHIGHI, A., Peyravi, F., and zadeh, A.,(2011), "A New Digital Signature Algorithm", 2009 International Conference on Machine Learning and Computing, IPCSIT vol.3 (2011), IACSIT Press, Singapore.

[19] Noroozi, E., DAUD, S., SABOUHI, A., and NAMADI, M.,(2012), "New Implementation of Hashing and Encoding in Digital Signature", 2012 IACSIT Hong Kong Conferences, IPCSIT vol. 29 IACSIT Press, Singapore.

[20] Rached, T., Mohsen, M., Noura, B., and Wajih, E.,(2012), "Low Power Elliptic Curve Digital Signature Design for Constrained Devices", International Journal of Security (IJS), Volume (6) : Issue (2).(1).

[21] Rjasko, M.,(2008), "Properties of Cryptographic Hash Functions", Cryptology Archive, ePrint, Report 2008/527.

[22] Shehory, O.,(1998), "Architectural Properties of Multi-Agent Systems", The Robotics Institute, Carnegie Mellon University - Pittsburgh, Pennsylvania 15213.

[23] Tonien, D., To, V., and Seberry E.,(2011), "A New Generic Digital Signature Algorithm", Groups, Complexity, Cryptology, 3 (2), 221-237.

[24] Brown, D.,(2009), "SEC 1: Elliptic Curve Cryptography", STANDARDS FOR EFFICIENT CRYPTOGRAPHY, Version 2.0.

[25] Wang, A.,(2001), "Using a Mobile, Agent-based Environment to support Cooperative Software Processes", M.Sc., Norwegian University for Science and Technology, Department of Computer and Information Science.

[26] Yang, Q.,(2008), "A Multi -agent Prototype System for Helping Medical Diagnosis", Thesis (Masters), Memorial University of Newfoundland.

[27] Zhang, C., Xi1, J., and Yang, X.,(2008), "An Architecture for Intelligent Collaborative Systems based on Multi-Agent", IEEE, 978-1-4244-1651, Official U.S. Government information about the Global Positioning System (GPS) and related topics.

[28] U.S. National Security Agency.,(2016), "Information Assurance Directorate", Commercial National Security Algorithm Suite and Quantum Computing FAQg.

[29] Singh, A., Juneja, D., and Sharma, A.,(2011), "Elliptical Curve Cryptography Based Security Engine for Multiagent Systems Operating in Semantic Cyberspace", International Journal of Research and Reviews in Computer Science (IJRRCS),Vol. 2, No. 2.

[30] Pal, V., Chowdary, P., and Gaur, P.,(2016), "A Review of Security Algorithms in Cloud Computing", International Journal Of Science & Technoledge (ISSN 2321 – 919X), Vol. 4, Issue. 10.

Firefly Algorithm Implementation Based on Arduino Microcontroller

Riyadh Zaghlool Mahmood
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
riyadh_1970@yahoo.com

Marwa Adeeb Al-jawaherry
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
marwaaljoharjy123@gmail.com

Abstract— There are many optimization algorithms; "Bio inspired metaheuristic algorithms" are one of the most important optimization algorithms. In this research, we intend to implement "Firefly Algorithm (FA)", which is one of the "Bio inspired metaheuristic algorithms" to optimize the finding operation of the maximum and minimum values of various mathematical equations based on Arduino microcontroller. The results are displayed on the GLCD, the following information is displayed: the number of the iteration (Itre), the minimum value (x), the maximum value (y) of variables in mathematical equivalents, the value of lightness (l), and finally the value of error (E).

Keywords: Optimization, Firefly Algorithm, minimum and maximum values, mathematical equations, Arduino mega2560.

I. INTRODUCTION

In most engineering and scientific problems, optimization is one of the most important ways to solve it, and through the continuous development in recent years many methods of optimization developed to optimize the solving of these problems. The most public methods are the metaheuristics methods [1].

At present one of the most common algorithms in global optimization problems is the nature-inspired metaheuristic algorithms", especially NP hard optimization. An example of those algorithms is the Swarm Optimization algorithm, developed in 1995 by Kennedy and Eberhart; these algorithms relied on the behavior of natural systems such as the bird schooling and fish. This algorithm was recently applied to find optimal solutions for many optimization applications [19].

The first source of inspiration for the design and development of many new optimization problems is the behavior of natural systems, such as ants systems, which is developed by observing the nature of ants system in nature, swarm intelligence is the behavior applied by these algorithms. It is therefore dependent on the interaction of individual entities and its social behavior is inspired from the behavior followed by insects [12].

The firefly algorithm was developed by the "Xin-She Yang", a firefly algorithm inspired by the behavior of fireflies in nature, two thousand firefly species is the estimated number of the their population. Most of these fireflies produce rhythmic and short flashes. Bioluminescence process generates flashing light of the fireflies. It may serve as warning signals or an element of court ship rituals [18].

In this research we choose to design and implement the firefly algorithm to find maximum and minimum values of mathematical equations, Arduino microcontroller was used to develop our proposed system and the system results were displayed on GLCD.

II. RELATED WORK

Bidar M. and Kanan H. R. [4] proposed an algorithm inspired from Firefly algorithm. The researchers intend to record the behavior of the all fireflies to recognize the weak ones, and enables them to update their locations by jumping to new locations in order to obtain find the solution, when the fireflies modified their locations that lead to modify the locations of whole population. The jumping operation increase probability of finding the optimal solutions, as so as increasing the performance of the proposed algorithm.

El-Sawy A. A. and et. al. [7] suggested a new approach that combines between two optimization techniques, "ACO and FFA". The propose approach was tested on many optimization problem such as benchmark problems, by applying this combining approach the researchers found that his performance was better than the performance of each approach when it is work alone.

Garsva G. and Danenas P. [10], their paper suggests new approach for linear classifier optimization method. Experimental results refer to the ability of proposed approach get competitive or better results compared to another similar approach. The linear classifier optimization approach can used to solve several classification problems with efficient solutions.

Asokan K. and Ashok Kumar R. [2], they propose an innovative optimization approach for defining bidding techniques, is shown as a stochastic optimization problem. The firefly algorithm introduced to this problem to optimize the search operation for best solution. By applying this approach the GENCOs profit maximizes in an effective way. Six suppliers was introduced to illustrate the main features of this approach, all results were displayed.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. FIRFLY ALGORITHM

In 2007, Firefly algorithm was used for the first time, [13]. It was used to optimize the Intelligence swarm algorithms. The method of this algorithm depends on the nature behavior of the firefly and the bioluminescent method for interaction between them [13] [15].

The difference in the value of light intensity is the value that is relied on by an objective function of an optimization problem. Depending on this value, fireflies update their locations as they move to the most attractive locations to reach the optimal solutions. Thus, light intensity that is related to the objective function is the characteristic of all fireflies [6].

A. Characteristics of Firefly Algorithm

Three basic rules were found for the Firefly algorithm, which rely on the main flashing characteristics of the behavior of living fireflies in nature. The rules were as follows:

1. All fireflies are "unisex" so fireflies will attract individual firefly.
2. The attractiveness is proportional to their brightness. The brighter fireflies attract other fireflies which has less bright. However, when the distance between two fireflies increase, the intensity should decrease
3. Fireflies move randomly, if the fireflies have a same brightness level.

By computing the value of the objective function, the firefly's brightness can be determined [16].

B. Functions of Firefly Algorithm

1. Attractiveness

Firefly's attractiveness function has its own form that can be illustrated in the decreasing function that described in equation 1.

"r is the distance between any two fireflies, r=0 is the initial attractiveness at r=0, and γ is an absorption coefficient which controls the decrease of the light intensity" [2].

$$\beta(r) = \beta_{0\, e^{(-\gamma r^m)}} \qquad with\, m \geq 1 \qquad (1)$$

2. Distance

If there are two fireflies i and j, the distance between them can be found in the following equation:

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \qquad (2)$$

"x_i, is the k-th element of the i -th firefly position within the search space, and d denotes the dimensionality of a problem" [9].

3. Movement

The following equation shows the movement of fireflies attracted by the most attractive fireflies [3]:

$$X_{i+1} = x_i + \beta_0 e^{(-\gamma r)2} (x_i - x_j)^2 + \alpha\, (rand - 0.5) \qquad (3)$$

"The second term is due to the attraction while the third term is the randomization with being the randomization parameter". Where "rand" was a random number generator, "rand" value was distributing in the range of [0, 1] [8].

The following pseudo-code form presents the firefly algorithm [6].

1. Algorithm's parameters initialization:

- Number of fireflies (n_f).
- $\beta 0$, γ, α
- Maximum number of generations (iterations, n_{itre}).

2. Define the objective function f(x), x = (x1. . ., xd)^T.

3. Generate initial population of fireflies xi (i = 1, 2 . . ., n). Light intensity of firefly I_i at x_i is determined by value of objective function f(x_i).

4. While k<n_{itre}

5. For i = 1:n

6. For j = 1:i

7. If (Ij> Ii) move firefly i towards firefly j in d-dimension according to Eq. (3); End if.

8. Obtain attractiveness, which varies with distance r according to Eq. (1).

9. Find new solutions and update light intensity

10. End for j.

11. End for i.

12. Rank the fireflies and find the current best

13. End while

14. Find the firefly with the highest light intensity.

The following equation represents the initial population of fireflies :

$$xi = LB + rand \cdot (UB - LB) \qquad (4)$$

Where LB and UB denotes the lower and the upper bounds of i-th firefly [6].

IV. ARDUINO MEGA2560

Arduino "is an open-source physical computing platform based on a simple I /O board", It takes the inputs of variety sensors or switches, and has the ability to control many devices and send different types of outputs such as lights, and other outputs, Arduino is therefore used to develop objects that need to interact with their external environment or with other objects, as they can interact with computer programs such as flash and processing [17]. However, Arduino can accomplish many projects on its own; and it has a special development environment for writing programs [5]. We use Arduino Mega 2650 to develop our system. See figure 1.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

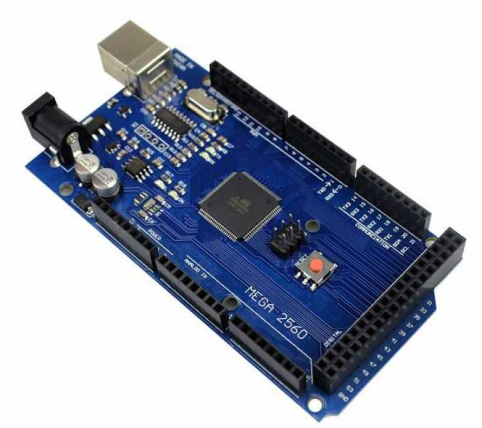


Figure 1: Arduino mega2560

Arduino designed for people with little technical and programming expertise, the use of Arduino allows these people to create a sophisticated model for project design and interactive artworks. People who have a strong technical background will be very easy for them to apply first steps with Arduino [14].

V. GLCD 192*64

A graphic LCD "liquid crystal display" is one of the electronic technologies that used in visual display, and also used in different gadgets and information output sources.

Through precise electronic signals, GLCD technology can employ manipulating tiny crystals of a contained liquid crystal solution to perform graphic display operations over a two dimensional screen.

LCD technology uses electron firing gun to produce a pixel based display over monitor screens, if traditional CRT "cathode ray tube" technology is compared with LCD technology; the latest technology is more successful [11] [20].

Figure 2: Graphic LCD.

We use Unified Modeling Language (UML) to develop our proposed system; use cases diagram, activity diagram and sequence diagram were used to analyze the system. See figures 3, 4 and 5.

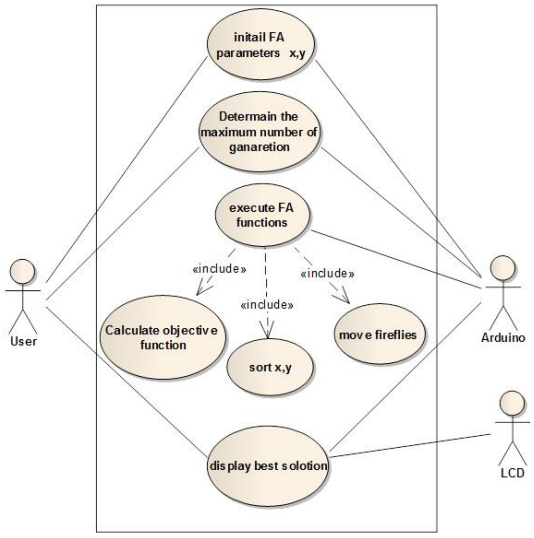


Figure 3: Use case diagram.

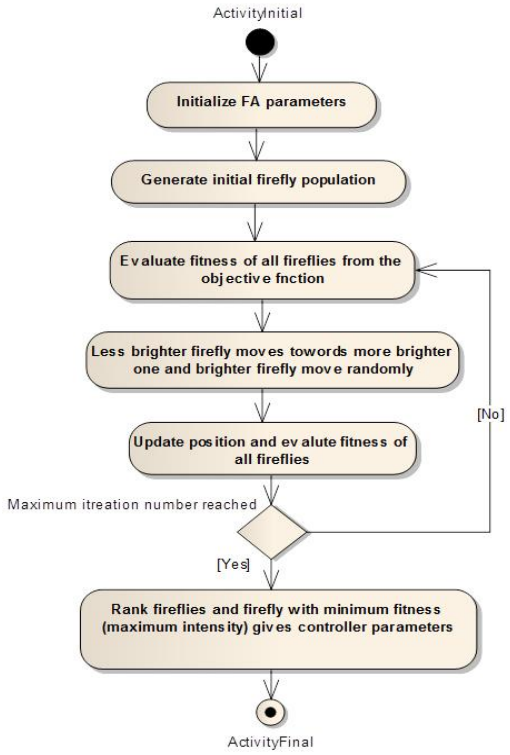


Figure 4: Activity diagram for firefly algorithm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

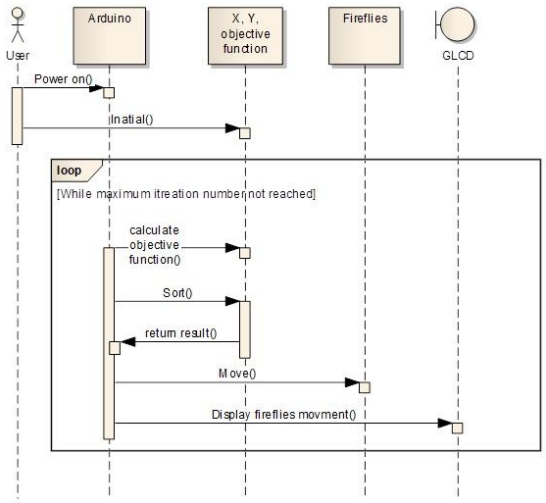


Figure 5: Sequence diagram.

VI. OPTIMIZATION TO FIND THE MINIMUM AND MAXIMUM VALUE OF VARIABLES IN MATHEMATICAL EQUATIONS USING FIREFLY ALGORITHM.

The firefly algorithm used to solve many optimization problems. For doing this we need to determine the objective function and the control parameters that can be decision variables, is given in equation 5:

$$\min, max \rightarrow f(x,y) = x + cy^2 - xy\,, (x,y) \in (-w,w)$$

Where c denote "any constant number", and w denote the "lower and the upper bounds of i-th firefly", the value of the x and y variables is choose randomly.

In our research we implemented firefly algorithm based on Arduino under windows XP or windows seven operating system and the result displayed at GLCD. In the implementation of any technique of metaheuristic techniques, the control parameters must be initialized. This also applies to the firefly algorithm, and it is very important to choose appropriate values for the control parameters to find the best solutions, the assigned values of control parameters are determining the performance of this method. Our selection of these parameters depends on a wide range of experimental results.

The control parameters displayed as following:

A. n_f: is the fireflies number, in all examples $n_f = 45$. We choose this value because when we set up n_f to a large number "more than 100 fireflies", the results of our experiments are not change greatly and the execution time was increase with no improvement at all.

B. n_{iter}: is iterations number, n_{iter} is another control parameter of the firefly algorithm which must be appointed to execute the algorithm until achieving the convergence of the minimization of the error. In order to find the global optima, the firefly algorithm was not need to large number of n_{iter}.

In all our experiments, n_{iter}= 50. We found that the value was a suitable,

When we increase the n_{iter} more than 50 iteration, the result does not improve.

C. β_0: The initialization value of attractiveness, as several suggestion for many optimization problems the value of β_0= 0.1. In the present study we take an above value, which give to us very good results.

D. γ: is the absorption coefficient, where $\gamma = 1$ in our paper, this value produce a convergence of the algorithm quickly.

E. μ: the value of potential coefficient, it can be assigned to any positive number. The value of μ = 0.1in our study.

F. α: the value of randomization parameter. This control parameter can be any number on the interval [-2.048, 2.048]. The randomization degree was determined by α value. The parameter α was so important because it was allowing to produce a new solutions, so as not to stuck in a local minimum. In our research α = 0.1, we choose this value in order to avoiding perturbations on the firefly.

First we must choose control parameters value. The firefly algorithm is performed iteratively until reached the number of iterations. To remove the stochastic effect and avoid premature convergence, 20 independent executions have been carried out. Then, the firefly with the best fitness value was selected as the optimal solution to the given problem.

VII. EXPERIMENTAL RESULTS

In this section we check the performance of our work, it has been tested with a large collection of examples, and the results were excellent in all cases. In this section we consider only one of these examples. These examples were selected to illustrate the variety of situations that could be applied using this method

The example in this paper is shown in Figures 1, 2 and 3. Three different figures are displayed: on the Figure 1, we show the first iteration that display the primary locations of the fireflies, on the figure 2,we show the eight iteration that display the new locations of the fireflies and on the figure 3, we show the twenty one iteration that display the fireflies are reaching to goal. As we say before our results will display on GLCD, this information is: the number of the iteration (Itre), the minimum value (x) and the maximum value (y) of variables in mathematical equations, the value of lightness (l), and finally the value of error (E).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 6: The first iteration that display the initial locations of the fireflies

Figure 7: The ninth iteration

Figure 8: Twenty one iteration that display the fireflies are

VIII. CONCLUSION

The firefly algorithm is an effective technique in solving global optimization problems. The firefly algorithm was used in this paper to find the minimum and maximum value of variables in mathematical equations using Arduino microcontroller.

The suggested approach depends on choosing the value of control parameters of firefly algorithm like: number of iterations, absorption coefficient, determination of the objective function, and population number of fireflies. Experimental results show that the results that were obtained is matching with desired results.

It is therefore possible to say that swarm Intelligence algorithms are highly efficient in solving optimization problems, including finding minimum and maximum value of variables in mathematical equations.

REFERENCES

[1] Abdullah A. and his colleagues, (2012), "A New Hybrid Firefly Algorithm for Complex and Nonlinear Problem", Springer, distributed computing and artificial intelligence, AISC 151, pp 1.

[2] Asokan K. and Ashok Kumar R. ,(2014), "Application of Firefly algorithm for solving Strategic bidding to maximize the Profit of IPPs in Electricity Market with Risk constraints", international journal of current engineering and technology, vol.4, no.1, pp 37-44.

[3] Benabid R. and his colleagues, (2014),"Application of Firefly Algorithm for Optimal Directional Over current Relays Coordination in the Presence of IFCL", I.J. intelligent systems and applications, pp 48.

[4] Bidar M. and Kanan H. R., (2013), "Jumper Firefly Algorithm", international conference on computer and knowledge engineering ICCKE , Qazvin, Iran, pp 278-282.

[5] Cytron technologies, (2011), "Fun and learning with Arduino projects ", Beginner guide, pp 4.

[6] Dhillon M. and Goyal S.,(2013), "PAPR Reduction in Multicarrier Modulations Using Firefly Algorithm", international journal of innovative research in computer and communication engineering vol. 1, issue 5, July , pp 1272.

[7] El-Sawy A. A. and his colleagues, (2012), "A Novel Hybrid Ant Colony Optimization and Firefly Algorithm for Solving Constrained Engineering Design Problems", journal of natural sciences and mathematics, qassim university, vol. 6, no. 1, pp 1-22.

[8] Farook S. and Raju P. S., (2014), "Metaheuristic Algorithms for Capacitor Sitingand Sizing to Improve Voltage Profile", International Electrical Engineering Journal (IEEJ), Vol. 5, No. 1, pp. 1211.

[9] Fister I. and his colleagues, (2014), "On the Randomized Firefly Algorithm", Springer International Publishing Switzerland, pp 34.

[10] Garsva G. and Danenas P., (2014), "Particle swarm optimization for linear support vector machines based classifier selection, Vilnius University, vol. 19, no. 1, pp 26–42.

[11] Jackson K. (2010), "Using a HD44780 Compatible LCD module with an Arduino microcontroller and Programming Environment", https://www.arduino.cc/en/Tutorial/LiquidCrystalDisplay.

[12] Kwiecien J.andFilipowicz B.,(2012), "Firefly algorithm in optimization of queuing systems", bulletin of the polish academy of sciences, technical sciences , vol. 60, no. 2, pp 1.

[13] Sarbazfard S. and Jafarian A., (2016), "A Hybrid Algorithm Based on Firefly Algorithm and Differential Evolution for Global Optimization",

[14] International journal of advanced computer science and applications, vol. 7, no. 6, pp 96.

[15] Schmidt M., (2011), "Arduino A Quick-Start Guide", North Carolina, Texas, Pragmatic Programmers, LLC, Second Edition.

[16] Wang G. and his colleagues, (2012), "A Modified Firefly Algorithm for UCAV Path Planning", international journal of hybrid information technology, vol. 5 issue 3, p123.

[17] Winkler F., (2007), "Arduino Workshop", pring,http://www.arduino.cc.

[18] Yang X. S., (2009), "Firefly algorithms for multimodal optimization", lecture notes in computer sciences 5792, pp 169.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[19] Yang X.S.,(2010), "Firefly Algorithm, L´evy Flights and Global Optimization", Springer Verlag London Limited, pp 1.
[20] www.vishay.com/docs/37364/37364.pdf.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Analysis of android bugs for mobile applications

Tawfeeq M. Tawfeeq Al-Flaih
Software Engineering Department
College of Computer Science and Mathematics
University of Mosul
Mosul, IRAQ
Tawfeek.mflaih@yahoo.com

***Abstract*— Open source Mobile applications have gained a lot of popularity in today's world. But most of these mobile applications tend to be buggy which may affect the user experience and thus they need quick bug fixes. For our research we have taken into account 10 applications from different domains. The aim is to study the bug reports of these applications and analyze them. Our objective for this research is to understand the life cycle of Android bugs and the relationship between the various domains and ratings with the number of bugs.**

Keywords: mobile applications, Android bug report, Google play store, bug fixing, bug report quality

I. INTRODUCTION

Mobile devices have become an important part of people's lives in recent years. Smartphone's have gone beyond their basic communication functions and now offer many features that in the past belonged solely to the domain of personal computers. As a result, companies have developed mobile versions of applications that were originally for other platforms. There is also a large quantity of applications developed specifically for mobile [13]. Multiple Apps stores were created by large companies to accommodate and manage their platforms Apps. On the other hand, and due to the wide spread of these mobile application, software repositories were used to maintain and share open source code for such applications. Software repositories such as source control, bug and, communication repositories are widely used in large software projects [7].

Application stores (e.g., Google Play, Apple App Store and BlackBerry App World.) have changed the traditional software development concept by providing their own platform for the rapid growth of mobile apps .in the past few years, mobile apps have exploded into a multi-billion dollar market and their popularity become hugely wide among consumers and developers. Mobile app downloads have risen from 7 billion in 2009 to more than 197 billion in 2017. In the same time, mobile apps numbers have also increased: Google Play now hosts over 28 million mobile apps [2][3].

In this paper we want to shed some light to understand the life cycle of open-source Android Apps bugs. To accomplish our goal we analyzed the bug reports of ten open-source Android Applications trying to understand the life cycle of these bugs. Furthermore, we are trying to measure the quality of these bug reports.

II. MOTIVATION

Recently, Android platform and its applications have gained tremendous popularity. The septal to entry in applications development and deployment has drop, due to easy distribution across application stores such as Apple App Store [8]. This means that's apps and app updates are subject to limited audit before deployment, and in this case there are many error-prone applications in the market and affecting user experience. Open source Mobile applications have gained a lot of popularity in today's world. But most of these mobile applications tend to be buggy which may affect the user experience and thus they need quick bug fixes. Most of open source Mobile applications have the bug report to gain feedback from users. User reports bug he have and describes some bug information. The bug will be opened and finally be closed. From "open' to "close", there is a life cycle of bug. Understanding the life cycle of bug can help us to reduce the bug occurrence.

Compared with iOS application, the Android applications will be run on many kinds of mobile drives. In this case, the cost of checking and fixing bugs for Android application will be more expensive than on iOS. For reducing this cost, the most effective method is that decreasing the number of bugs before the application released. We hope we can find some properties about Android application through analyzing the bugs and bug report.

III. RELATED WORK

There is a lot of research performed related to life cycle of bugs in the android applications. Bhattacharya P. et. al. in the paper "An Empirical Analysis of Bug reports and Bug fixing in open source android apps" performed an empirical analysis so as to understand the bug fixing process in the Android platform and Android based applications. In order to perform their research, they selected 24 popular android applications. They selected apps depending on certain metrics and analyzed the bug fix processes. This included the bug fix time, bug categories, bug priorities and also the interest of the users and developers to fix the bugs. On comparing the life cycle of bugs on Google Code and Bugzilla they found that lack of certain bug report attributes affects the bug fix process. They investigated the categories of security bugs in Android applications. On conduction of the analysis, they found that even though the contributor activity in the projects is high, the involvement of developers is less. Also, triaging bugs is still a problem even though the bug reports are of high quality. They

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

observed that the non security bugs required less time to fix even though the quality of security bug reports was better.

The MSR challenge provides platform for the researchers to add their mining tools and approaches to the challenge. There is a research done in the android platform for analyzing bugs and finding some interesting facts of those reports. This research is performed by Shihab E. et al. in the paper "Mining Challenge 2012: The Android Platform". The work is performed on the change data and bug report data of the android platform which has been extracted from GIT repository and android bug tracker. They selected sub-projects for change data from the android those are Kernel/linux, kernel/omap, kernel/tegra, Kernel/Samsung, kernel/qcmu, kernel/experim-enttal, platform/frameworks/base, platform/external, Bluetooth/ bluez. In the change data analysis the result states that the numbers of authors are more than the number of committers, which shows GIT has fewer contributors than committers so as to fix the issues. In a similar way for the bug report they selected 10 different components. Those are Market, Docs and Build, User, Web and System, GfxMedia, device, media, Google dalvik, tools, applications, platform and no component. The result set the average fix time for bug found is 2.34 months, most of bugs were not assigned to any of the particular component, the committers commit on the bug report only once during the project and 99% of the bugs are of medium priority and also it has a length of average 189 words.

Syer et al. performed a study on comparing the mobile applications with different desktop application. They considered two aspects for comparison, the size of the code base and the time to fix the defects. For conducting their study they considered 15 popular open source android applications and 5 different desktop applications. They found that there is a large difference between the mobile apps and desktop apps in some respects, while in some respects they are similar. They found that the core developers in mobile apps are very small as compared to desktop applications. Thus it is necessary to pay attention to mobile development now by keeping aside the desktop applications. In our research, we are going to study the life cycle of bugs in open source android applications.

IV. RESEARCH QUESTIONS

In order to conduct our study, we have identified the following research questions. Our objective of this research is to answer these questions:

1. How can the quality of bug reports help the contributors fix the bugs sooner?

2. What is the relation between domain of Application and number of Bugs?

3. What is the relation between the rating of Application and number of Bugs? Our methodology to answer these research questions is described in the next section.

V. METHODOLOGY-STUDY DATA

A. Selection Criteria:

The mobile applications we selected in this project are open source applications. There are millions of mobile applications in the market today. However, only 10 mobile applications are needed. So the selection criteria are narrowed by selecting android applications. The reason is its number of options available to select an application, the popularity of applications people using, these applications are available free of cost and main important reason is availability of its own bug repository with some of its applications.

The android mobile applications are downloaded from Google Play store. There are two categories of applications available in the play store. Those are free and paid. As name implies the applications can be downloaded at free of cost and paid applications can be downloaded by paying for it. The advantage of play store is, it provides 26 categories to choose the application. Moreover, the play store provides a detail like category, number of downloads, number of people rated it and also some time it provides link to Git repository. It is important to remember that not all the free applications of the play store comes with the GIT repository.

The Git repository provides all the details necessary for the bugs to analyze. The complete life cycle of the bug can be observed. The bugs from the initial release to the present releases can be found. The open bug count, closed count, data and time they were reported and fixed, and contributors and commenter's details can also be studied from here.

The table 1 provides the details of the 10 mobile applications selected for the project. The details are its category, the number of downloads, number of people rated and wrote the review, total number of releases and the bug count which is sum of the open bugs and closed bugs from the first release of the application.

We use some tools as CUEZILLA tool to measures the quality of new bug reports [4].

TABLE I. APPLICATIONS DATA

Name	Category	Downloads	Ratings	Releases	Bugs Count
Zxing: Barcode Scanner	Utility	100,000,000 – 500,000,000	704060	16	372
FBReaderJ	Education	10,000,000 – 50,000,000	128429	306	325
Wordpress	Editor	1,000,000 - 5,000,000	63185	69	131
Keepassdroid	Security	1,000,000 - 5,000,000	28904	110	317
Ifixit	Utility	500,000 - 1,000,000	5760	28	251
Simon Tatham's Puzzles	Game	100,000 - 500,000	31469	56	230
Car Cast	Multimedia	100,000 - 500,000	1240	77	121
BetterBattery Stats	System	100,000 - 500,000	7,986	139	601
AnkiDroid	Education	1,000,000 - 5,000,000	18,166	389	745
XBMC	Multimedia	100,000 - 500,000	281	81	343

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

RQ1: How can the quality of bug reports help the contributors fix the bugs sooner?

Our first research question was to understand how the quality of bug reports will help the contributors and developers to fix the bugs easily and quickly. In order to answer this question, we have taken into account different characteristics of bug reports. These characteristics include the length of description of the bug in the bug reports and number of keywords found in the description [1][5]. They keywords that we have considered are version, component, security, vulnerability, attack, failure, error, crash, buffer overflow, buffer overrun, question, problem, invalid, and incorrect. For every bug in the bug report for each application, used a script to find out the length of the description of the bug. Also we wrote a script which took the input as the above mentioned keywords and found them in the bug descriptions. The descriptions having highest number of keywords along with sufficient description length were chosen. The bug descriptions which were too lengthy and did not have high count of the keywords were ignored. Also the bug descriptions which were very short in length but had large number of keywords were discarded. Further, we calculated an average of description length and the average of number of keywords for each application.

The table 2 below shows an example of how each bug from every application was analyzed to find the description and number of keywords and the corresponding time spend to fix the bug.

TABLE II. BUG REPORT QUALITY

App	Bugs ID	Bug Title	Start Date	End Date	Time	Length of Description	Number of Keywords
AnkiDroid	105	Allow users to change AnkiDroid directory if current one is invalid	2/3/2015	2/15/2015	12D	190	3
FBReaderJ	219	Fatal exception in BookDownloaderService	1/10/2013	Open	N/A	171	3
Zxing	308	Possible ReedSolomon deoding problem	2/18/2014	2/20/2014	2D	253	4
CarCast	71	Review if debug mode is needed on release builds	10/26/2012	7/1/2014	978D	30	3
ifixitAndroid	106	SSL errors on Android 2.2	9/9/2013	9/10/2013	1D	313	4
KeepassDroid	39	2nd try…	10/21/2010	10/24/2010	3D	68	1
sgtpuzzles	9	Build Failed	2/19/2015	Open	N/A	79	3
WordPress	104	Bugfix - iploading post thumbanails	3/16/2013	3/16/2013	3min	86	2

Result:

We obtained the following results as described in the table 3 below. On careful analysis, we found that the time span required to fix the bugs was less for the bugs which had good description length along with large number of keywords. The keywords and description made it easy for the developers and contributors to understand the bugs and get them fixed as early as possible. As we can see for Zxing application the average length is 60 and average number of keywords for the bug report is 112 so the average fix time is less 10 days. Similarly for Simon puzzle application the average length is 46 and in proportion to the length, average keywords is 40 so time span is 6 days.

TABLE III. AVERAGE LENGTH, AVERAGE NO. OF KEYWORD AND NO. OF DAYS FOR BUGS IN EACH APPLICATION

Name	Total No of bugs	Avg Length	Avg Keywords No.	Avg Time Span in days
Zxing: Barcode Scanner	372	60	112	10
FBReaderJ	325	48	52	16
Wordpress	131	18	16	99
Keepassdroid	317	36	20	59
Ifixit	251	57	26	112
Simon Tatham's Puzzles	230	46	40	6
Car Cast	121	43	6	139
BetterBatteryStats	601	55	87	15
AnkiDroid	745	42	54	3
XBMC	343	52	61	4

RQ2: What is the relation between domain of Application and number of Bugs?

The second research question is focus on the relation between domain of application and the number of bugs. In our research, we choose four domains as our research objectives. We select four open source applications for each domain. In the same time, all of the applications we choose have substantially the same ratings. That means those applications have the same evaluation. Then we calculate the bug count for each application, and compare them based on the same domain.

Result:

The charts below show the relation between the application domain and the number of the bugs. As shown in the figures (1) are no clear variation between results number of bugs are very close between the domains. One of the apps in the education domain (ankidroid) has more bugs in comparisons with other domain but this is may not be because the domain bug may because the application itself or the nature of the team who developed this application.

RQ3: What is the relation between the rating of Application and number of Bugs?

The last research question to answer is relation between the number of ratings and bug count. In Google play store the people who downloaded it rate the application from 1 to 5 stars that is, from average to very good. Along with

ratings of the stars, the people write their review on the application use. The rating is sum of the people rated application and people wrote reviews for it. The bugs count is from the GIT repository. The analysis is made for each selected 10 applications.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

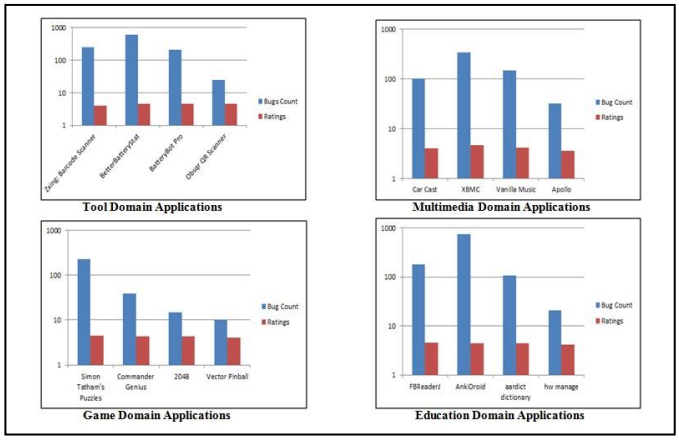


Figure 1. Graphs for rating of each domain

Result:

The results is found by analyzing the figure (2) below, which shows the relation between the ratings of the application and bugs count for each individual application. As we can see from figure 2, the rating and bugs count are inversely proportional. Higher the rating for an application lower is the bug count. In 10 applications, all the 9 application supports the conclusion other than one application XBMC, which is a multimedia application similar to VLC player. The reason for this is the number of downloads 100,000 - 500,000. This Application also has very less feature compared to other media players which makes its less popular and more buggy. This download number reveal that the people are less interested in using it so the less number of ratings 271. Hence the greater number of bugs counts.

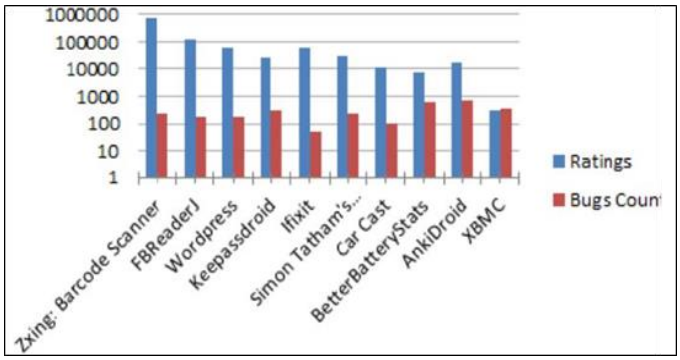


Figure 2. Relation between rating of application and numbers of Bugs

B. *ZXING Analysis:*

During the initial phases of the data collecting and analysis, ZXING application raised many question because of its high download and rating numbers. Also, we notice that its bugs count relatively low. Therefore, we decided to go further analyzing this application, and seek some answers and explanation for these numbers. To accomplish that goal, we extracted the end users reviews from ZXING application page in Google play store and classify them into feature requests and bug reports. Then, we tried to see if the users are satisfied with application and not asking for many features.

Analysis Approach:

After extracting the reviews from the Google Play Store, to answer our questions we needed to classify them in feature requests and bug reports. To achieve this, we wrote an algorithm that splits a text into sentences, normalizes them, and compares them with a set of linguistic rules to find if it matches any of them. We used two set of rules for our classification. One is based on the linguistic rules defined by Iacob et al. for feature requests [9], the other one is based on the linguistic rules they defined for bug reports [10]. We adapted the syntax of the rules to work with OpenNLP [12], the API we used for part-of-speech tagging. To classify issues, we also modified some rules, because the way people talk when reporting an issue is somewhat different than the way they talk when leaving a review at an app store. Table (4) shows some examples of linguistic rules for identifying feature requests and an example text that would be a match for each. Table (5) shows some examples of linguistic rules for identifying bug reports and an example text that would be a match for each.

TABLE IV. EXAMPLES OF RULES TO IDENTIFY FEATURE REQUESTS

Rule	Text match
Would be <adjective> if	It would be great if
Would <adverb> like to <verb>	Would really like to see
Needs option to	Needs options to share posts

TABLE V. EXAMPLES OF RULES TO IDENTIFY BUG REPORT

Rule	Text match
<adverb> annoying	Incredibly annoying
Won't <verb>	Files won't open
Keeps on crashing	Reader keeps on crashing

To create an algorithm that classified the reviews based on those rules, we used Lingpipe and Opennlp. We started by splitting the review in sentences, using Lingpipe [11] to recognize end of sentence tokens. Lingpipe is a toolkit for processing text using computational linguistics. After the review was split into sentences, we normalized each sentence, replacing common misspelled words and abbreviations. After, we used OpenNLP [12] to tag the sentence. OpenNLP is a machine learning based toolkit for the processing of natural language text. We used it to tag each word in the review sentence as a part of speech, e.g. for the text "it would be great"

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the tagger would generate "<personal_ pronoun> <modal> <verb> <adjective>".

We ran the algorithm twice for each review we extracted. The first time it compares the review sentence with the linguistic rules defined for feature requests, and if it matched one of the rules, it classified the sentence as a feature request. If a sentence of a review, then the review is marked as not feature request. The second time it did the same, but instead of comparing the sentence with the rules for feature requests, it compared them to the rules for bug reports. After the reviews in the database were classified, we counted the bug report and feature requests for the applications.

VI. DISCUSSION

Base on the results of three questions, we can clearly find out the quality of bug report affect the speed of fixing the bug. The keywords can make developers easily to understand and locate the bug. The length of bug report and the time for fixing is a negative correlation. This is easy to understand that the more information the developers get, the faster the bug can be fixed.

Also, the differences of domain generally have relation with quality of mobile application. Base on the same rating for each domain, the game mobile application have less number of bugs than other three domains. The one reason we think is that user have more patience for other three domains than for game. That means the mobile game developers have to pay more attention to reduce the incidence of bug. Through analyzing the rating of application and the number of bug, we found out the applications having higher ratings have less bug count. The bug can reduce user experience, thereby decrease the ratings.

VII. THREATS TO VALIDITY

The research would have been given better results if more number of applications were taken into consideration. For this study we have considered only 10 applications which is a very limited number. Also for our research question 2, we considered only 4 applications for each domain. If more number of applications were considered we could have got different results. In Addition, all the applications are written in the same language (Java). We did not explores the code of these apps such as the number of classes, the number of developers, and the experience of developers of these application, to get better results we supposed to select the apps that are close in the number of line of codes or the number of developers so that the domain will not be affected by the code.

When testing our review classification algorithm for Zxing analysis, we used our personal judgment to decide if the algorithm was right or not for identifying a feature request or a bug report. Therefore, the accuracy measures of the algorithm are biased.

VIII. CONCLUSION

In this paper, we analyzed some open source mobile applications, and get the relations between domain, rating, and the quality of bug report. Through analyzing the life cycle of bug, we realize that the users' behavior also affect the quality of mobile application. In this case, it proved that the importance of bug report. We also realize that the difference of user's patience for different domain. This also decides that there are different decisions of testing and fixing bug in the different domains. For the future work, we can increase the number of mobile applications which are analyzed to get richer data set. In the same time, we want to make clear for the correlation between length of bug report and the time of fixing. Is there a crest in their correlation? In other words, does the too much information in bug report effects the developers' understanding for the bug? We hope our research can provide some enlighten for who also analyze this area.

IX. REFERENCES

[1] Bhattacharya, P., Ulanova, L., Neamtiu, I., & Koduru, S. C. (2013, March). "An empirical analysis of bug reports and bug fixing in open source android apps". In Software Maintenance and Reengineering (CSMR), 2013 17th European Conference on (pp. 133-143). IEEE.

[2] Shihab, E., Kamei, Y., & Bhattacharya, P. (2012, June). "Mining challenge 2012: The android platform". In Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (pp. 112-115). IEEE Press.

[3] Syer, M. D., Nagappan, M., Hassan, A. E., & Adams, B. (2013, November). "Revisiting prior empirical findings for mobile apps: an empirical case study on the 15 most popular open-source Android apps". In Conf. Center for Advanced Studies on Collaborative Research (CASCON 13), IBM, 2013, pp. 283–297.

[4] Bettenburg, N., Just, S., Schröter, A., Weiss, C., Premraj, R., & Zimmermann, T. (2008, November). "What makes a good bug report?". In Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering (pp. 308-318). ACM.

[5] Asaduzzaman, M., Bullock, M. C., Roy, C. K., & Schneider, K. A. (2012, June). "Bug introducing changes: A case study with Android". In Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (pp. 116-119). IEEE Press.

[6] Martie, L., Palepu, V. K., Sajnani, H., & Lopes, C. (2012, June)."Trendy bugs: Topic trends in the android bug reports". In Proceedings of the 9th IEEE Working Conference on Mining Software Repositories (pp. 120-123). IEEE Press.

[7] Jongyindee, A., Ohira, M., Ihara, A., & Matsumoto, K, I. (2011, November). "Good or Bad Committers? A Case Study of Committers' Cautiousness and the Consequences on the Bug Fixing Process in the Eclipse Project". In Software Measurement, 2011 Joint Conference of the 21st Int'l Workshop on and 6th Int'l Conference on Software Process and Product Measurement (IWSMMENSURA) (pp. 116-125). IEEE.

[8] Sinha, V. S., Mani, S., & Gupta, M. (2012, June). "Mince: Mining change history of android project". In Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on (pp. 132-135). IEEE.

[9] Iacob, C., and Harrison, R. (2013, May) "Retrieving and analyzing mobile apps feature requests from online reviews". in Proceedings of the 10th Working Conference on Mining Software Repositories, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, pp. 41–44.

[10] Iacob, C., Harrison, R., and Faily, S. (2014) "Online reviews as first class artifacts in mobile app development" in Mobile Computing, Applications, and Services. Springer, pp. 47–53.

[11] alias i, LingPipe, 2003U″ 2011. [Online]. Available: http://alias-i.com/lingpipe/

[12] T. A. S. Foundation, OpenNLP, 2010. [Online]. Available: https://opennlp.apache.org/

[13] Jiang, H., Yang, H., Qin, S., Su, Z., Zhang, J., Yan, J.,(2017, Dec) "Detecting Energy Bugs in Android Apps Using Static Analysis" In: Duan Z., Ong L. (eds.) Formal Methods and Software Engineering. ICFEM 2017. Lecture Notes in Computer Science; Vol. 10610: pp 192-208 : DOI: 10.1007/978-3-319-68690-5_12.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Enhanced Pipeline for Improving Reliability using Evolution and Optimization Technique

Dr Narasimha Rao Yamarthi[1], Mr Adinew Belay[2], Mr Abdisa Lechisa[3], P.Narasimha Rao[4].
Professor, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
HOD, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
Lecturer, Department of Computer Science, School of Computing & Informatics, Mizan Tepi University, Ethiopia.
narasimha.yamarthi@gmail.com, adinewb2@gmail.com, abdisalechisa@gmail.com,pnr555@gmail.com

Abstract

Objective: In advanced digital systems the propagation delay plays a vital role to optimize the performance of an individual processing element. In the present paper the advantages and flaws of various pipelines are discussed. In the present paper the performance evaluation of different pipelines is done in terms of various parameters like timing delay, throughput, and average delay. These factors are very important in achieving parallel computing in fast processors.

Methods/Analysis: In the present paper the proposed pipeline is compared with Traditional Pipeline, Wave Pipeline, and mesochronous Pipeline. In all the cases the throughput, Timing delay, and average time delay are compared and proved that the proposed method has produced more effective parameters. All the observations are made at 4-stage pipeline. The design analysis is done with the simulation software Proteus. The accurate data wave reliability is tested in Proteus. The propagation delay is illustrated with the help of Electronic Work Bench. The readings are distinguished at different data frequency rates like 100, 500 and 1000MHz.

Findings: It is observed that a four stage proposed pipeline has good throughput when compared with other pipeline clock schemes. Based on the observations the wave pipeline is superior to any other method in terms of through put and data reliability. The proposed method achieved slight improvement when compared with wave pipeline. The data reliability is good in proposed method at different frequency stages.

Novelty/Improvements: To achieve parallelism in advanced processors, pipeline technique is the best method proposed by many architectural designers. In real time operating systems the pipeline helps in message passing and fetching in due time. But there are many design and operational factors need to be considered in achieving high performance. The Propagation delay is one of the important factor need to consider in pipeline design.

Keywords: Parallelism, Pipeline, Clock Scheme, Propagation Delay, Efficiency, Throughput, Reliability

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

1. Introduction

In the present paper a data pipeline is discussed in terms of its performance based upon the clock scheme. There are different clock schemes effectively acting on internal latches of pipeline. Different timing constraints are indicated in the previous work held in the present area[1,2]. The digital pipeline play very important role in processing the data with memory and I/O devices. Pipelines plays most vital role to avoid bottle neck. Pipelined processors can be clocked a fast clock rate and thus can have reduced cycle times (more cycles/second by a fast clock) than un-pipelined implementations of the same processor[3]. In traditional pipeline the flow of data that is inputting the data, intermediate process, and outputting the data from stage to stage is controlled by common clock cycle. The process between all stages is controlled by common clock cycle. The intermediate latches are used between stages to hold the intermediate process results. The stages are basic combinational circuit. Intermediate latches can also be used for delay balancing in data path[4].

In the present paper a linear pipeline with new clock scheme is discussed with multiple parameters which are set to prove the efficiency. For a general pipeline the time delay is denoted with τ.

Where, $\tau = \tau_m + \tau_l$ ---------- (1)

A linear pipeline with k stages uses k cycles to fill up the pipeline and n-1 cycles are needed to complete the remaining n-1 tasks. In the present design a static pipeline of unifunction is discussed with their performance. Here the pipeline is designed to operate at different pipeline bandwidths. The bandwidth represents the number of bits processed per unit time. Here the performance of the pipeline is measured in two factors[5].

Pipeline Efficiency: The efficiency of linear pipeline is measured by percentage of busy time-space spans over the total time-space spans over the total time-space span, which equals the sum of all busy and idle time-space spans [5]. Let *n*, *k*, and τbe the number of tasks, number of pipeline stages and the clock period of linear pipeline respectively, then the pipeline efficiency is defined by

$$\eta = \frac{n}{k + (n-1)} \quad \text{---------- (2)}$$

Throughput: The number of results that can be completed by a pipeline per unit time is called its throughput. This rate reflects the computing power of a pipeline. Throughput can be defined as shown,

$$w = \frac{n}{\tau} \quad \text{---------- (3)}$$

Average Delay in a stage is,

$$\tau_{avg} = \frac{T}{k} \quad \text{---------- (4)}$$

The time required to finish i^{th} instruction in a pipeline computer is T_i

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$T_i = (N_i + k - 1).\frac{T}{\kappa} \quad \text{---------- (5)}$$

Where,

τ_l= the delay of each interface latch
τ_m = the delay through longest logic path
k= the number of stages in a functional pipe
T= the total pipeline delay in one instruction execution
n= the number of instructions contained in a task
N_i= the length of vector operands used in the i^{th} instruction
W= the throughput of the pipeline computer
T_i=the time required to finish i^{th} instruction in a pipeline computer
η= the efficiency of a pipeline computer

In most of the cases the pipeline performance efficiency depends on the effective clock signalling. There are some mostly used clocking schemes are already playing an important role in steering the pipeline performance, such as synchronous, Asynchronous, Mesynchronous, and Plesynchronous clock schemes. In synchronous clocking maximum power consumption occurs due to global data. Higher clock speed is required and less clock periods will be used for computations. In Asynchronous clocking, the drawback is that the hardware and signalling overhead involved in the local communication and in any timing constrains that are required by particular choices of signalling protocols [6,7]. Plesynchronous interconnect only occurs in distributed systems like long distance communications. The data can be duplicated if the transmit frequency is slower than the receive frequency. These problems can be overcome with the new clock scheme. The above factors are observed by considering four different pipeline techniques, out of which one is the new method.

i) Conventional pipeline: In conventional pipeline system a single clock pulse is applied to manage the data transmission through the registers in the pipeline. But it will create a clock skew in thepipeline which will decrease the data speed from one stage together stage [8,9,10,11].

ii) Wave Pipeline: Smaller clock periods are achieved in wave pipelining[12] byreducing the maximum propagation delay (τ_m) by splitting the stages into number of stages[13]. The width of the clock pulse will be approximately equal to the difference between maximum and minimum propagation logic path delays between pipeline stages.

iii) Mesynchronous Pipelining: The propagation delay is reduced and the clock synchronization is controlled by introducing a delay element in the path of clock signal of Mesynchronous pipelining [14]. The delay element is almost equal to the logic path delay between pipeline stages.

2. Enhanced Method

A four stage pipeline is constructed to analyse pipeline operation and process the data. A four stage pipeline is proposed because; an *n* stage pipeline performs *n* faster operation in any type processor. Intermediate latches are used between stages to hold the intermediate process.

In order to achieve proper capturing of data at the output proper clock timing must be done between stages. The timing requirements must be met between clock and data edges at

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the inputs to the output. The clock period must be such that the output data is latched after the latest data has arrived at the outputs and before the earliest data from the next clock period arrives.

In the present method the logic gates at individual stages create simple delay in producing the clock to the next stage. Until the logic gates identify the next binary bit from previous stage it will not allow the clock generator to pass the next clock pulse to the next stage of the circuit as shown in Figure 1.

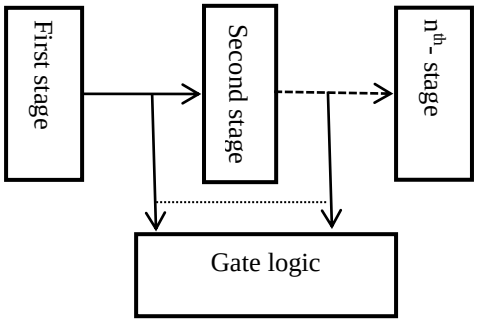


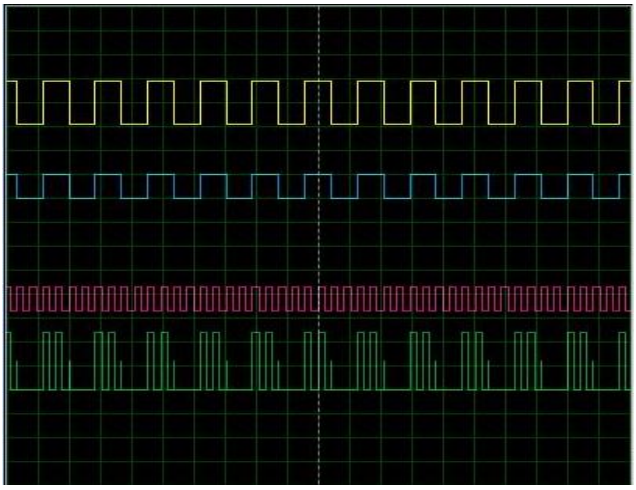
Figure 1. Proposed Method with new clock scheme

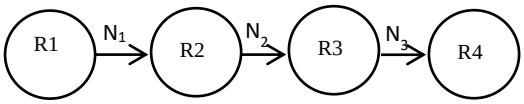
In the present paper this method is compared with other traditional, wave and mesochronous pipelines in terms of timing delay, throughput and average delay.

In traditional pipeline the data propagation is not accurate when compared with other optimized pipeline techniques. In synchronous pipeline a common clock pulse is unable to synchronize all stages at different frequencies.

In traditional pipeline the data propagation is not accurate when compared with other optimized pipeline techniques. In synchronous pipeline a common clock pulse is unable to synchronize all stages at different frequencies. It results loss of data bits in transmission. In asynchronous traditional pipeline as the stages increases the circuit complexity also increases and hence propagation delay. The propagation delay increases and results data latency in prior stages. Hence, the reliability is deprived in traditional pipeline. There may be a chance of fault event due to miss match between clock rate and data rate. This causes failures at individual phases [15]. If the failure rate is repeatedly occurring as shown in Figure 2 the reliability of the pipeline present system will be low. In Figure 2 the third and fourth pulses are input and output pulses respectively. In the present traditional pipeline system, there is a failure case identified at each third and fourth pulse due to fault occurrence because of synchronization problem between clock and data frequency. If the fault detection at any stage is N_k, where k is number of the stage, then the faults at individual stages can be represented with the help of model [16,] as shown in Figure 3. R_1 and R_2 are the reliability factors of stage 1 and stage 2 respectively, and reliability can be represented upto n stage like R_n.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500





In Wave pipeline and Mesochronous pipeline based system the reliability is improved when compared with traditional method. These two are the optimized clocking methods used to reduce the failures. In wave pipeline likewise faults arouse due to propagation difference between longest and minimum path difference between stages. This leads irregular data propagation failures as shown in Figure 4 some data loss is observed in fourth clock pulse. These failures are observed due to clock skews due to (D_{max}~D_{min}) of wave pipeline [13,17]. The faults at individual stages can be represented with the help of Jelinski-Moranda model as shown in Figure 5.

In the proposed method the data propagation is monitored at every stage with special control circuitry to enhance the accuracy. Even when the clock frequency is not synchronized the pipeline stages will control the previous data through gate logic as shown in Figure 1. It is observed that data pulses are propagated accurately when compared with other existing methods. An accurate data waves are observed at different timing rates as shown in Figure 6.

To maintain higher performance of the pipeline predefined accuracy levels and different failure rates are defined. For materializing a mathematic equation and for evaluating reliability following assumptions are made.

The system contains N homogeneous stages and failure density depends on number stages and it is exponentially distributed for i=*n.* In the present system, λ_c is the failure rate due to the clock skew, λ_p is the failure rate due to the delay difference of logic path between stages, λ_g is the failure rate due to the gate logic control. Gate logic control is used to control the data wave propagation to the next stages. R(t) is the reliability of the pipeline at individual stage at predefined sample size. The reliability function is evaluated as,

$$R(t) = \sum_{i=0}^{n} P_i(t) \qquad \text{---------- (6)}$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The probability of failure rate distribution due to λ_c,λ_p, and λ_g, is P(t),

$$P_i(t) = \left(\lambda_i\left(N-i+1\right)+\lambda_p+\lambda_g+\lambda_c\right)e^{\left(-\lambda_i\left((N-i+1)+\lambda_p+\lambda_g+\lambda_c\right)t_i\right)} \quad \text{---------- (7)}$$

$$where, i = 1,2,3,...N$$

The reliability between first and second stage is R_{p12} and reliability of data propagation between second and third stage will be R_{p23} and so on, and R_{p12}, R_{p23} are in series.

Then the total reliability is given by=

$$R_{p12} = \left[1-\prod_{i=1}^{2}(1-R_i)\right]$$
$$R_{p23} = \left[1-\prod_{i=2}^{3}(1-R_i)\right] \quad \text{---------- (8)}$$

$$R(t) = R_{p12}R_{p23} \quad \text{---------- (9)}$$

For n stage pipeline the overall Reliability is,

$$R(t) = \sum_{i=1}^{n} R_{pi(i+1)} \quad \text{---------- (10)}$$

3. Result Analysis

The circuits for respective pipelines are constructed in Proteus and Electronic Work Bench. The reliability of the circuit design is tested, modified and analysed in Proteus.

The results are obtained and analysed with Electronic Work Bench to analyse data throughput. The results are obtained as shown in Table 1 at different frequencies.

A four stage pipeline is constructed and analysed for n=4. The reliability model is designed and mathematical formulas are evaluated in section 2. An accurate data waves are observed and so system design is reliable.

4. Conclusion

The parameters are observed on a four stage pipeline and assumed number of tasks is equal to 4. The performance is analysed at different data rates starting from 5Hz to 1GHz. In this paper readings at higher data rates are represented in the table and graph. It is observed that the new method showing optimistic results in timing delay, throughput and average delay when compared with other three methods. But in case of wave pipeline it is observed better results than traditional and Mesynchronous pipeline, with fewer logic gates. And still the new method need to be observe by cascading higher stages. The reliability of the proposed method is found high by considering failure rate λ_g, and other failure rates.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

5. Reference

[1] Gay CT. Timing constraints for wave pipelinedsystems. IEEE transactions on Computer aided design of integrated circuits. 1994 Aug, 13 (8), pp. 987-1004.
[2] Kumar NS, Reddy DV. Effect of Interrupt Logic on Delay Balancing Circuit. International Journal of Computer Applications. Published by Foundation of Computer Science New York, USA. 2011 Aug, 27(4), pp. 26-30
[3] TOKHEIM RL. Schaum's Outline of Theory and Problems of Computer Architecture Copyright © The McGraw-Hill Companies Inc. Indian Special Edition. 2009, pp. 1-17.
[4] Chang CH, Davidson ES. Delay balancing using Latches. 2000, pp. 1-9.
[5] Hwang K, Briggs FA. Computer Architecture and parallel processing. McGraw-hill International Editions, Computer Science Series, 1st edition. 1990.
[6] Kumar NS, Reddy DVRK, Sridevi R, Sridevi V. Studies and a Method to Minimize and Control the Jitter in Optical Based Communication System. IJACSA. 2013, 4 (4), pp. 45-51.
[7] Huang HY, Cheng KH, Wang JS, Chu YH, Wu CY. LOW-voltage Low-Power CMOS True Single-Phase Clocking scheme with Locally Asynchronous Logic Circuits. IEEE Xplore. 1995, 3, pp. 1572-1575.
[8] Shrivishwanath D, Raajan NR, Gopinath L. Flip Flop and Double Substitution Based Textual Encryption. Indian Journal of Science and Technology. 2014 Apr, 7 (4), pp. 52-55.
[9] Vignesh B. Pipelined Quadratic Equation Based Novel Multiplication Method for Cryptographic Applications. Indian Journal of Science and Technology. 2014 Apr, 7 (4), pp. 34-39.
[10] Kumar NS, Reddy SVRK. A New Method to Enhance Performance of Digital Frequency Measurement and Minimize the Clock Skew. 2011, 11 (10), pp. 2421-2425.
[11] Elavarasi R, Kumar PKS. An FPGA Based Regenerative Braking System of Electric Vehicle Driven by BLDC Motor. Indian Journal of Science and Technology. 2014 Nov, 7 (7), pp. 1-5.
[12] Gray CT. Timing constraints for wave pipelined systems. IEEE transactions on Computer aided design of integrated circuits. 1994 Aug, 13 (8), pp. 984-1004.
[13] Burleson WP, Ciesielski M, Klass F. Wave pipelining: a tutorial and research survey. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 1998 Sep, 6 (3), pp. 464 -474.
[14] Tatapudi SB. A Mesychronous high performance digital systems. 2006 May, 53 (5), pp. 1-19.
[15] Kumar NS, Reddy DVRK. Measurement system for wide range frequency at nonlinear devices. FCICS Published by Taylor and Francis group. 2015, pp. 143-146.
[16] Joe H, Reid N. On the software reliability models of Jelinski-Moranda and Littlewood. IEEE Transactions on Reliability. R. 1985, 34 (3), pp. 216-218.
[17] Rao YN, Raju SV. Enhancing Data Fetching Rates with Parallel Pipeline. International Journal of Computer Applications. 2014, 85, pp. 31-34.

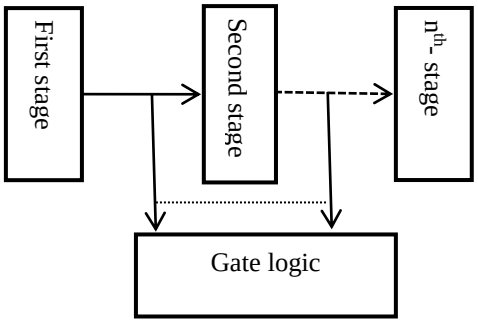


Figure 1. Proposed Method with new clock scheme

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

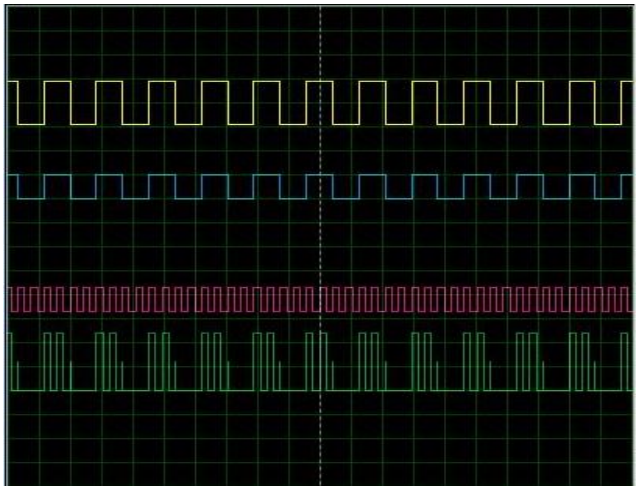


Figure 2. Presence of Failure in traditional pipeline

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

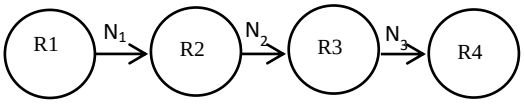


Figure 3. Fault assumptions of Jelinski-Moranda model for traditional pipeline

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

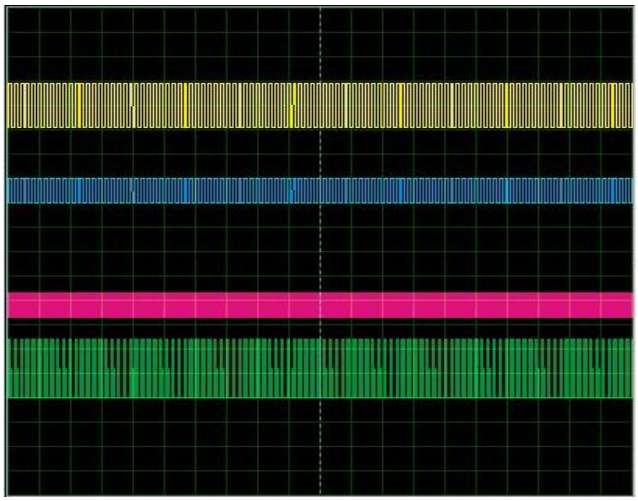


Figure 4. Data Propagation through Wave pipeline Clock scheme

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

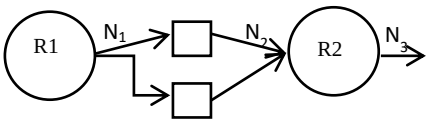


Figure 5. Fault assumptions of Jelinski-Moranda model for Wave Pipeline

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

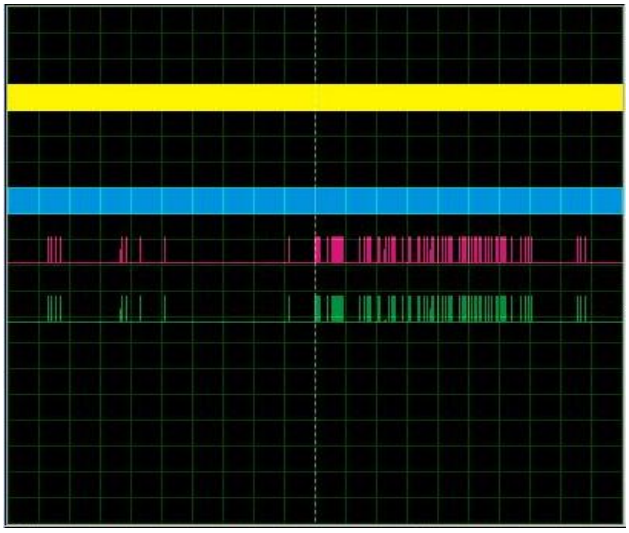


Figure 6. Data Propagation through New method

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1. Results of different four stage pipelines

Data Frequency in MHz	Traditional pipeline			Wave Pipeline			Mesochronous Pipeline			Proposed Pipeline		
	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τavg) in µsecs	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τ_{avg}) in µsecs	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τ_{avg}) in µsecs	Timing delay (τ) in µsecs	Throughput (*w*) in MHz	Average Delay (τ_{avg}) in µsecs
1000	1.7	2.3	0.425	1.8	2.2	0.45	1.7	2.3	0.425	1.4	2.86	0.35
500	1.78	2.25	0.445	1.78	2.25	0.445	1.67	2.4	0.418	1.44	2.77	0.36
100	1.89	2.12	0.473	1.89	2.12	0.473	1.78	2.25	0.445	1.62	2.47	0.405

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Efficient Voice Based Person Identification System for Secured Communication

Dr.Shakila Basheer

Lecturer in Information system Department

King Khalid university

ABHA

Saudi arabia

Abstract

Secured Communication is essential due to scalability due to increase number of devices and drastically growing number of people involved in communication. In this paper a voice comparison based communication authentication mechanism is used for providing secured communication. This voice based authentication is used in two different applications like people communication and data retrieval. Before going to speak with people in online their information and their voice is compared and verified from the database and permission will be granted. Similarly according to the voice they can retrieve the data from the data base, where it provides data integrity. Both applications comprise a number of stages such as: (i) Voice, Voice to Text input, (II). Voice Comparison and Pattern Matching. Finally (III). Permission Granted and Data Retrieval (DR) as the output. In order to improve the accuracy and relevancy the proposed data retrieval system, it uses an indexing method called Bag of Words (BOW). BOW is like an index-table which can be referred to store, compare and retrieve the information speedily and accurately. Index-table utilization in DRS improves the accuracy with minimized computational complexity. The proposed DRS is simulated in DOTNET software and the results are compared with the existing system results in order to evaluate the performance.

Keywords: Information Retrieval System, Data Mining, Bag of Words, Data Base Maintenance.

Introduction

In general IR is an activity is used by a few people for library management, paralegals and the digital library searching system. The world is growing with lots of changes were more than million number of people are using IR in everyday life like email, web searching. After sometime the IR system is used for information access and traditional searching in databases such as, searching an order, searching a product, searching a document from a digital library and so on. It is well known that the IR retrieves data from unstructured databases. The term “unstructured data” means the data is not clear, semantically overt and the format of the data is undefined. Simply can say that it is opposite to structured data (example: DBMS, RDBMS), but in real-time there is no data are not truly unstructured.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Searching information, images, documents and files are created based on the visual appearance and the properties of the data, document and images. Information retrieval is a challenging problem where it has been received a considerable attention from most of the researchers in various fields of image processing, data mining, information retrieval and computer vision and multimedia systems. The growth of web technology brings a drastic increase in data usage published in the recent decades, which has been a great challenge to develop efficient information retrieval systems to help all the users in IR systems. Traditional IR models such as: vector space model [16], classical probabilistic IR models [15] and language modeling approaches [13] are used for query based document retrieval and works independently. Web search engines are used for entity based retrieval [14, 12] used for commercial purpose. An entity based web document retrieval [9-11] are used in the earlier research works to provide a better semantic based document searching. Searching, information retrieval, content based information retrieval systems are still getting urgent demand in the web applications [17,18]. The retrieval system concentrates on features as important for information extraction. Most of the paper follows the feature based IR on content based image retrieval systems [19-21]. Some of the IR systems used to transform in order to decompose and represent various resolutions, various sizes and various amounts of information [22-23]. Wavelet transform have been successfully applied to image Denoising [24], image compression [25] and texture analysis [26]. In [27] the authors propose a new CBIR system using color and texture features. In this paper texture features are extracted Euclidean distance measure to obtain the similarity measurement between a query text and text in the database. In [28] wavelet basis was used to characterize each query image and also to maximize the retrieval performance in a training data set. To make DRS is more efficient, DRS is not constructed based on all the entities. It is query independent. For each voice query the index is selected and then the related data are selected from different location. One the index is matched, and then DRS decides the location of the data and the entities of the index-data from the database. In this paper the information retrieval system is developed using index searching and pattern matching methodologies. To do index searching BOW is used. The contribution of the proposed DRS work is:

- Voice (input)
- Creating BOW
- Voice Matching and Pattern Matching
- Communication Permission granted and Data Retrieval (output)

Proposed Model

The proposed model clearly says about the entire functionality of the proposed DRS and it is shown in Figure-1. Any physically challenged people one who are not able to operate the keyboard can use this application. In this paper, it is assumed that the application is developed for online shopping. The user can say about the product in mic then the voice is converted into text. The converted text is taken as a keyword for pattern matching in the product database. During the pattern matching keyword is verified with the BOW in order to check the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

product availability. If the keyword is available in BOW then the other relevant information about the product is taken from the database, converted into voice, and play back to the user. It is an advanced application can be used in handheld devices also.

In user communication, initially the numbers of users are registered with their voice. The voice is the keyword for comparison, whereas before coming to communicate in online both end user has to be verified by the voice. If the present users' (ready for communication) voice is matched with the stored DB voice then they are permitted for communication and they can proceed. The this functionality is depicted in Figure-1.

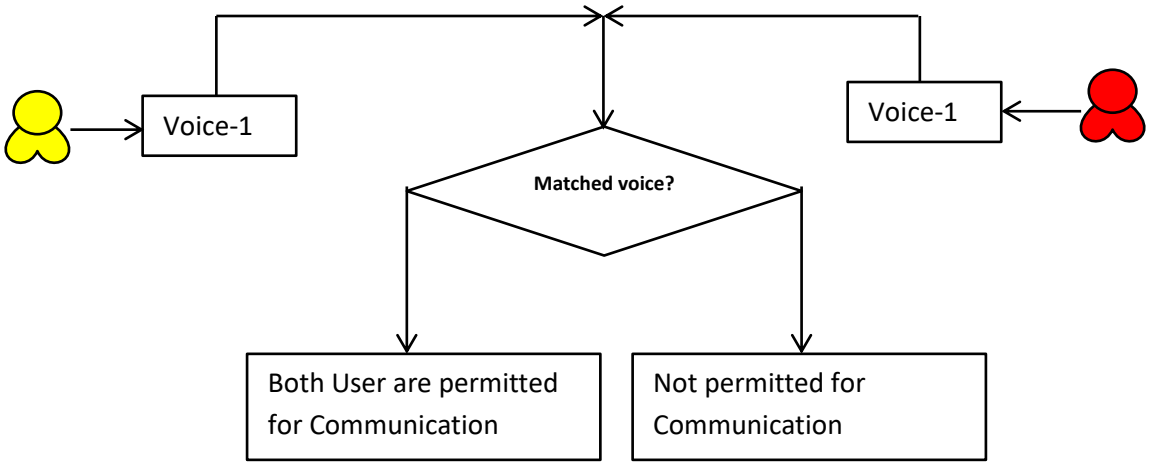


Figure-1: Application-1 [Secured User Communication]

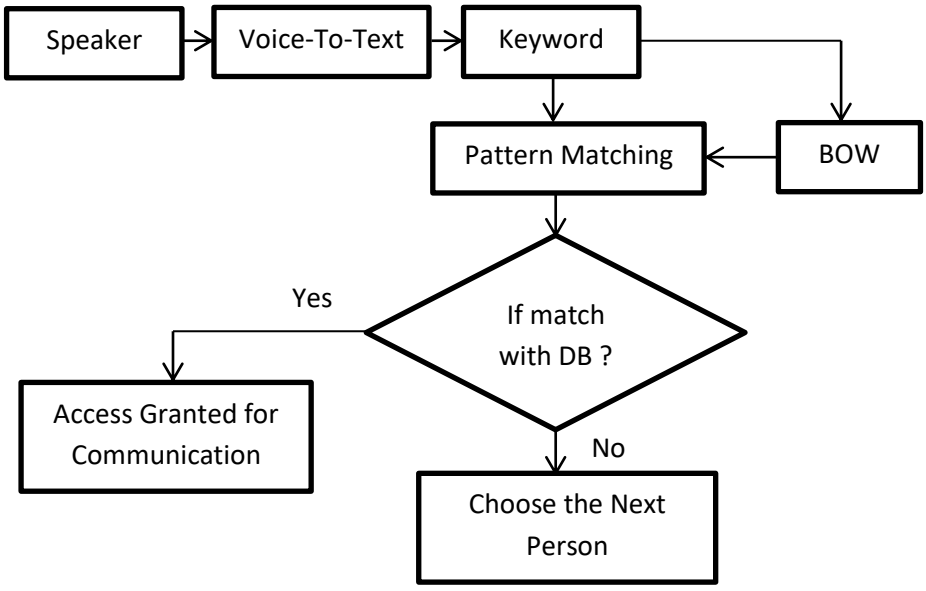


Figure-2: Application-2 [Secured Data Retrieval]

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Bag-of-Words

One of the most common methodologies to obtain the entire data is by visual words and it can be applied as text indexing and retrieval scheme. The index is created from any one of the feature of the data stored in the DB before persisting newly in the DB. It can be called as Bag-of Words or bag of feature model. Some of the static terms are taken from the data and it is maintained as a catalog (BOW). This catalog is compared with the database data for retrieving the specific data matched with the catalog. The data retrieval using keywords can predict maximum relevant based data and it satisfies the customer.

In this proposed DRS, whenever a new product detail is entered into the database, any one of the data feature is added as index word into the BOW. It needs not be a numeric or character data type and it can arrange the entire BOW automatic while inserting a new index. This automatically arranging of index words helps to compare and retrieve the relevant data speedily and accurately without computational complexity. For example: when a new data d_n is inserted into the database D, one of the feature from the feature set $f_i = \{f_1, f_2, \ldots, f_n\}$ is stored into BOW.

Field-1	**Field-2**	**…….**	**Field-i**	**……**	**Field-n**
D_1f-1	D_1f-12	…….	D_1f-i	……	D_1f-n
:	:	:	:	:	:
:	:	:	:	:	:
D_nf-1	D_nf-2	…….	D_nf-i	…….	D_nf-n

BOW
D_1f-i
:
:
D_nf-i

Figure-3: BOW Creation

Each field of the data is considered as separate features and any one of the field is stored into BOW. In an image retrieval system BOW is created automatically using LABELME tool. But in case of alphanumeric data the feature has chosen as keyword manually by the developer according to the convenient. Figure-3 shows the way of BOW creation and it can be used to compare the product availability in the database.

The data classification and retrieval is based on the BOW index, where BOW is the structured features taken from all the trained data inserted in the database. The word stored in the BOW belongs to the same class and it is behaving like a codebook used to cluster and classify the entire dataset. The words of all dictionaries represent frequent structures of all form types. Each word type is represented by a feature vector. The structural features of a form $\mathbf{s_j}$ are

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

calculated and are assigned to the cluster center $\mathbf{w_i}$ (word) with the smallest (Euclidian) distance mini $\|\mathbf{s_j}\text{-}\mathbf{W_i}\|$. This distance is used to fetch the matching BOW for the voice into text (keyword).

Voice –To-Text

A portion of the DRS system is programmed to recognize the speech (voice), and convert into text using speech synthesization mechanism available in the system library. The inbuilt speech recognition engine is instantiated initially, then the defined grammar is loaded in order to recognize the phrases. Adding grammar is used to identify the grammar-name. Each time the grammar is loaded dynamically in order to update the new BOW inserted. This updating can be obtained by the recognizer update method. In this paper the DRS listens to the user whether any speech data is entered into the system. The speech recognition engine is already loaded with the predefined trained text in the background. Each time speech made one line of text is displayed at a time in the system. The main advantage of this system is it will wait for a small interval in order to avoid congestion and proceed with the next BOW. If the speech is understandable by the speech engine then it keeps idle and waits for the next speech and it won't create any software breakup.

The speech to text is an application where it does translate words into text as much as possible due to various countries' accent variation. Other than the DRS, this voice to text conversion is used in healthcare, traffic systems, military, telephony and education systems. It is mainly focused for people with dis-abilities. This paper follows a fuzzy logic based Speech Recognition of Linguistic Content method [1]. In this method a word in a language, speaks in different accents, different speeds of pronunciation and with different emphasis. For example, the word "vector" of the English language will be spoken by an American as "vektor", with curtness at the 'c' and at the 't', while a Britisher will speak it as "vectorr", with emphasis on the 'c' and a slight repetition on the 'r'. Similarly, a Russian will speak this word as "vecthor", with softness on the 't'. However, the word remains the same, that is, "vector", with slight variations with respect to different accents, speeds of pronunciation and emphasis.

Thus, a single word can be represented as a fuzzy set. However, a word is too specific so as to fit into a generic model of speech recognition. To have a more general model, the fuzzification of phonemes is more appropriate. This model is therefore applied to spoken sentences. One fuzzy set is based on accents, the second on the speeds of pronunciation and the third on emphasis. The use of this method will be especially for speech-to-text conversion, by filtering out the unnecessary paralinguistic information from the spoken sentences.

Pattern Matching

In this paper the main idea is to search from right to left in the pattern. With this scheme, searching is faster than average. In order to do this the Boyer-Moore (BM) algorithm positions the pattern over the leftmost characters in the text and attempts to match it from right to left. If no mismatch occurs, then the pattern has been found. Otherwise, the algorithm computes a shift;

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

that is, an amount by which the pattern is moved to the right before a new matching attempt is undertaken. The shift can be computed using two heuristics: the match heuristic and the occurrence heuristic. The *match* heuristic is obtained by noting that when the pattern is moved to the right, it must

1. Match ***all*** the characters previously matched, and

2. To bring a different character to the position in the text that caused the mismatch.

The last condition is mentioned in the Boyer-Moore paper [3], but was introduced into the algorithm by Knuth et al. [2]. Following the later reference, we call the original shift table $\boldsymbol{dd}$, and the improved version $\widehat{\boldsymbol{dd}}$. The formal definitions are

$$\widehat{\boldsymbol{dd}}[\boldsymbol{j}] = \mathbf{min}\{\boldsymbol{s} + \boldsymbol{m} - \boldsymbol{j} | \boldsymbol{s} \geq \mathbf{1}\ \boldsymbol{and}\ ((\boldsymbol{s} \geq \boldsymbol{i}\ \boldsymbol{or}\ \boldsymbol{pattern}[\boldsymbol{i} - \boldsymbol{s}] = \boldsymbol{pattern}[\boldsymbol{i}])\ \ \boldsymbol{for}\ \boldsymbol{j} < \boldsymbol{i} \leq \boldsymbol{m})\}$$

$$\boldsymbol{for}\ \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{m}; \boldsymbol{and}$$

$$\widehat{\boldsymbol{dd}}[\boldsymbol{j}] = \mathbf{min}\{\boldsymbol{s} + \boldsymbol{m} - \boldsymbol{j} | \boldsymbol{s} \geq \mathbf{1}\ \boldsymbol{and}\ ((\boldsymbol{s} \geq \boldsymbol{j}\ \boldsymbol{or}\ \boldsymbol{pattern}[\boldsymbol{j} - \boldsymbol{s}] \neq \boldsymbol{pattern}[\boldsymbol{j}])\ \boldsymbol{and}\ ((\boldsymbol{s} \geq \boldsymbol{i}\ \boldsymbol{or}\ \boldsymbol{pattern}[\boldsymbol{i} - \boldsymbol{s}] = \boldsymbol{pattern}[\boldsymbol{i}])\ \ \boldsymbol{for}\ \boldsymbol{j} < \boldsymbol{i} \leq \boldsymbol{m})\}$$

The $\widehat{\boldsymbol{dd}}$ table for the pattern $\boldsymbol{abracadabra}$ is

dd	**a**	**b**	**r**	**a**	**c**	**a**	**d**	**a**	**b**	**r**	**a**
$\widehat{\boldsymbol{dd}}$ **[j]**	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**13**	**12**	**4**	**1**

The occurrence heuristic is obtained by noting that we must align the position in the text that caused the mismatch with the first character of the pattern that matches it. Formally calling this table *d*, we have

***d*[*x*] = min{*s*|*s* = *m* or (0 ≤ *s* < m *and* pattern *[*m - s*]* = x*)}**

for every symbol ***x*** in the alphabet. This methodology is used to compare the voice converted text with BOW and with the database. If the pattern matches the database, then the voice based reply is produced to the physically challenged people. The voice is produced by converting the relevant record information obtained from the database and convert into voice.

Text-To-Voice

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Text-to-speech synthesis takes place in several steps. The TTS systems get a text as input, which it first must analyze and then transform into a phonetic description. Then in a further step it generates the prosody. From the information now available, it can produce a speech signal. The structure of the text-to-speech synthesizer can be broken down into major modules:

Natural Language Processing (NLP) module: It produces a phonetic transcription of the text read, together with prosody.

- Digital Signal Processing (DSP) module: It transforms the symbolic information it receives from NLP into audible and intelligible speech. The major operations of the NLP module are as follows:

- Text Analysis: First the text is segmented into tokens. The token-to-word conversion creates the orthographic form of the token. For the token "Mr" the orthographic form "Mister" is formed by expansion, the token "12" gets the orthographic form "twelve" and "1997" is transformed to "nineteen ninety seven".

- Application of Pronunciation Rules: After the text analysis has been completed, pronunciation rules can be applied. Letters cannot be transformed 1:1 into phonemes because the correspondence is not always parallel. In certain environments, a single letter can correspond to either no phoneme (for example, "h" in "caught") or several phoneme ("m" in "Maximum"). In addition, several letters can correspond to a single phoneme ("ch" in "rich"). There are two strategies to determine pronunciation:

In dictionary-based solution with morphological components, as many morphemes (words) as possible are stored in a dictionary. Full forms are generated by means of inflection, derivation and composition rules. Alternatively, a full form dictionary is used in which all possible word forms are stored. Pronunciation rules determine the pronunciation of words not found in the dictionary.

In a rule based solution, pronunciation rules are generated from the phonological knowledge of dictionaries. Only words whose pronunciation is a complete exception are included in the dictionary. The two applications differ significantly in the size of their dictionaries. The dictionary-based solution is many times larger than the rules-based solution's dictionary of exception. However, dictionary-based solutions can be more exact than rule-based solution if they have a large enough phonetic dictionary available.

Whenever a voice input into DRS it is taken as the query for searching the relevant product from the database. Query enriches expansion is a general strategy used in text retrieval, which is directly adapted to the BOW model in all kinds of data retrieval. In this project the query expansion is simply taken as index searching with BOW and pattern matching with the database. There are various query methods are available like Transitive Closure Expansion (TCE) [4], and Additive Query Expansion (AQE) [5]. In this paper the TCE is used for query

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

processing system. Initially the query word (voice to text) is compared with the index where each visual word has an index indicating that the entire data is available in the database or not. This paper doesn't calculate the score value defining the similarity [6], since the keyword is unique. Using the above text to speech conversion the voice reply is generated and play with the user. The entire functionality of the proposed DRS is given in the form of algorithms, it can be coded in any computer programming language and the efficiency can be evaluated.

Algorithm_DRS (string product)

{

Input: voice, product data, initial BOW;

Output: voice

Description:

1. user speech in mic
2. Voice is converted into text
3. Apply a pattern matching algorithm
4. Search text into BOW
5. If (text exists in BOW) then search in DB
6. Voice (" product details"); // all the fields from the matched field is converted into voice
7. Else
8. Voice ("product not available");
9. End
10. If any product insertion then
11. field-i insert into BOW

}

Experimental Setup

The functionality of the proposed DRS is programmed in DOTNET 2010 software and the results are produced. There are 25 systems are installed in a laboratory in order to evaluate the system performance. In all, the system DOTNET software and the IRS module is installed. The proposed DRS are programmed, experimented in DOTNET software and the results are given below to analyze the performance. One among the systems is assumed as the server and the database is installed. The database is a lexical dictionary which consists of a collection of data in the form of rows. Each row consists of various numbers of columns which is not having appeared like a table. Another system is assumed as a middleware, having BOW table, which consists of a set of all inserted index keywords. Whenever a voice input entry to the system it

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

refers the BOW first and then comes to the database server, which reduces the computational complexity.

In order to experiment the proposed DRS, a product dataset is taken from [8] and experimented. 100 different products are stored in the database. It is assumed that the most of the product names are known by the user and it is online shopping. Some of the product name with some more relevant information about the product is shown in Table-1. Product code, product name are the two main features mostly used for searching the product information speedily in the entire database. Instead of concentrating all modules of online shopping, it is simply coming to know the product availability and product price with other relevant information about the product. The database consists of 15 fields in the table were on our paper only 5 fields are taken as important information to verify the DRS performance. In common product-code is used as searching indexes, but here due to voice mining, product name is used as searching indexes.

Table-1: Product Information

productCode	productName	productLine	quantityInStock	buyPrice	productDescription
S10_1678	1969 Harley Davidson Ultimate Chopper	Motorcycles	7933	48.81	This product is good and u can get world service
S10_1949	Alpine Renault 1300	Classic Cars	7305	98.58	Turnable front wheels; steering function; detailed interior; detailed engine; opening hood; opening trunk; opening doors; and detailed chassis
S10_2016	1996 Moto Guzzi 1100i	Motorcycles	6625	68.99	detailed engine, working steering, working suspension, two leather seats, luggage rack, dual exhaust pipes, small saddle bag located on handle bars, two-tone paint with chrome accents, superior die-cast detail , rotating wheels , working kick stand

There are 100 data is stored in the table where during searching computational time is spent only 100 comparisons and data fetching. For **an N number** of comparisons the computation time taken is **2N+2.** The following figures show that the efficiency of the proposed DRS in terms of accuracy, timeliness and response generations. In order to evaluate the performance, the number of data used in the database table is changed and verified. The number of data is changed from 100 to 1000 and the performance is compared.

In this paper the user provides their input as voice through multimedia input device. The voice is recorded and recognized by the speech engine installed in the system and it is converted into text. The voice recognition is a big process if the Voice-accent is understood by the speech

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

engine then it converts the voice into text. In this process, the number of voices is recognized accurately for the voice input given into the DRS. In order to evaluate the voice recognition accuracy by the DRS, the number of voice input is increased and the recognition rate is calculated. The number of voice input may be changed from 25 to 250. Each round of experiments the number of voice input is increased by 25. Out of the input voice, the number of voices recognized by the DRS system is calculated and shown in Figure-3. Still Google-Voice play is also finding difficulties in terms of voice recognition. In the proposed DRS system the recognition rate is better and it is increased according to the number of voice input increases. The recognition rate is proportionally increased, according to the number of voice inputs getting increased. After successful recognition, the voice is converted into text (it is taken as a keyword) for comparison with the BOW. If the keyword matched with the BOW index, then directly compared with the database in order to process the pattern matching.

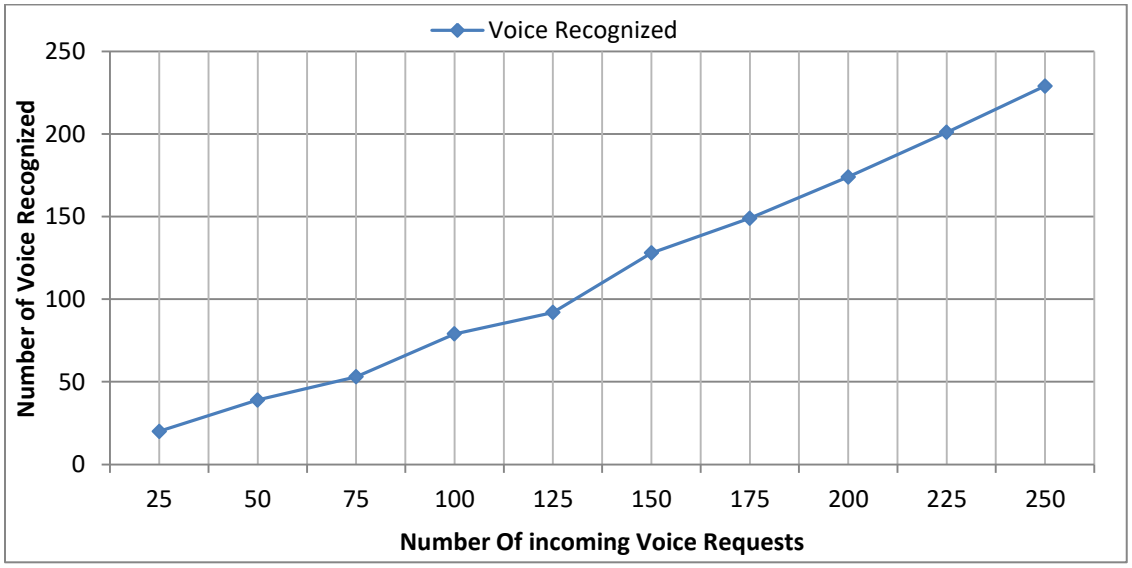


Figure-4: Number of Voice Inputs Recognized Vs. Number of Voice Inputs

If the pattern matched, then the relevant record data are fetched from the data row then converted into voice again. This text-to-voice conversion is played to the user who passed the voice input. According to the number of voice input processed, the number of voice reply is calculated and the quality of the DRS is verified. The number of voice reply against the number of voices is shown in Figure-5. Figure-5 says that the voice reply is increased according to the number of input voice. It is clear that after index matching the reply can be generated according to the pattern availability. The reply may be about the product or it is a message saying that particular product is not available and since there should be a compulsory voice reply for each voice input if it matched with the index. The execution process is preceded when the index is matched, else it is dropped executing the next process. Hence the proposed DRS reduces the computational complexity.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Also Figure-5 shows that the number of voice reply is merely equal to the number voice input given into the proposed DRS. It cannot be concluded that the pattern matching will be performed if the keyword matched with the BOW index due to the product may not be available. The pattern matching algorithms used in this paper find the distance between the possible patterns obtained from the DB with the input pattern. If the distance is merely equal to zero, then the pattern is matched, else it is not matched. According to the pattern matching algorithm, the accuracy is calculated and shown in Figure-6. The percentage of pattern matching is merely equal to the percentage of index matching. From this figure, it is clear that the number of pattern matching is lesser than the number of index matching. After the index matching successful the appropriate pattern may not available in the database and it affects the pattern matching accuracy. It cannot be concluded that the accuracy of the DRS is less. In this paper the accuracy of the entire IR system can be taken as the average of both index matching and pattern matching.

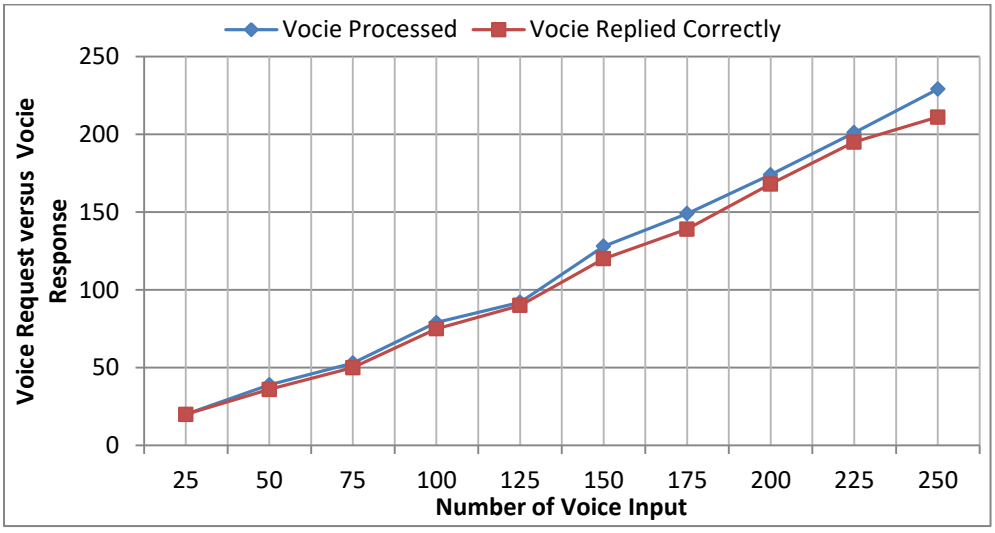


Figure-5: Number of Voice Input vs. Number of Voice Output

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

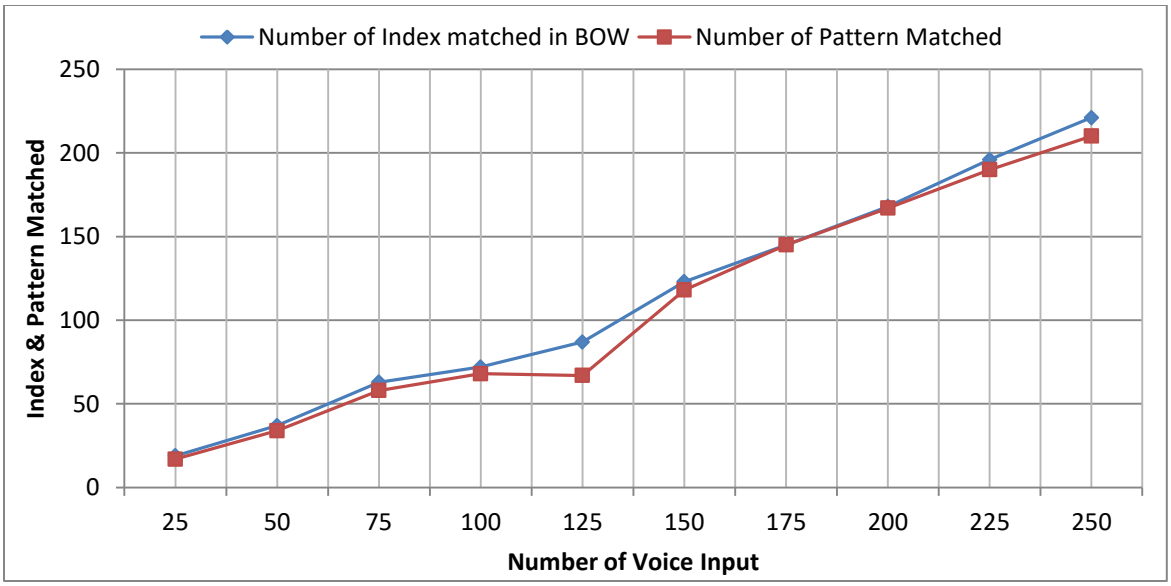


Figure-6: Voice Input Matched with Index and Matched With Pattern Comparison

The computational complexity refers the number of statements in the program to be executed in the compiler and the time taken to compile. The number of statements in the program decides the compilation time and the compilation time taken by the proposed DRS is shown in Figure-7. The figure shows that the computational time is less and it increases, according to the number of inputs increased. It means that for 100 numbers of data it takes only 4 seconds to make the entire process of DRS.

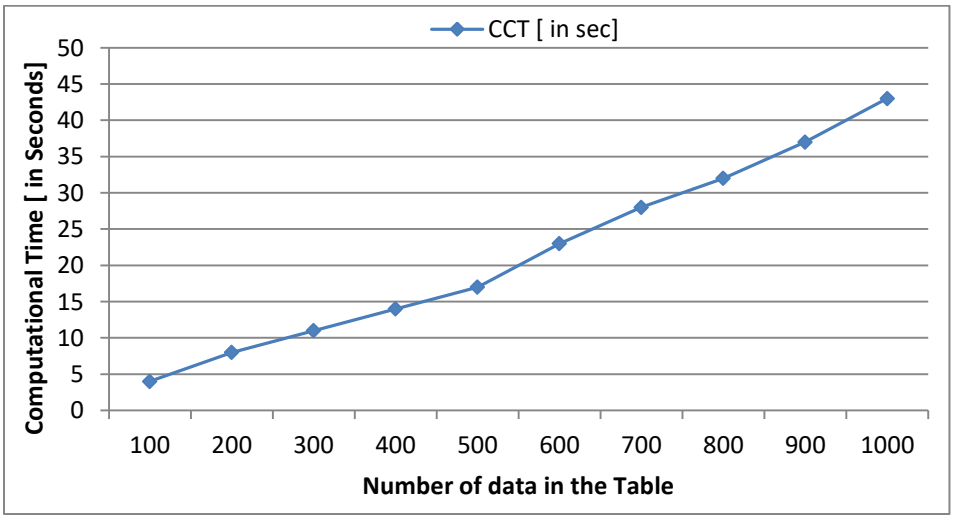


Figure-7: Computational Time In Terms of Data Size

Also the efficiency can be calculated according to the number of response generation against number of input queries. The number of query response against the number of input queries is shown in Figure-9. DRS proved that the number of pattern matching is not depending on the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

number of index matching completely. It depends on the index matching and the data availability. This figure shows the number of voice reply (response) provided to the user against the query input. The voice reply is gradually increased according to the number voice query applied. The accent and the data availability determine the accuracy of the pattern matching and voice reply accuracy.

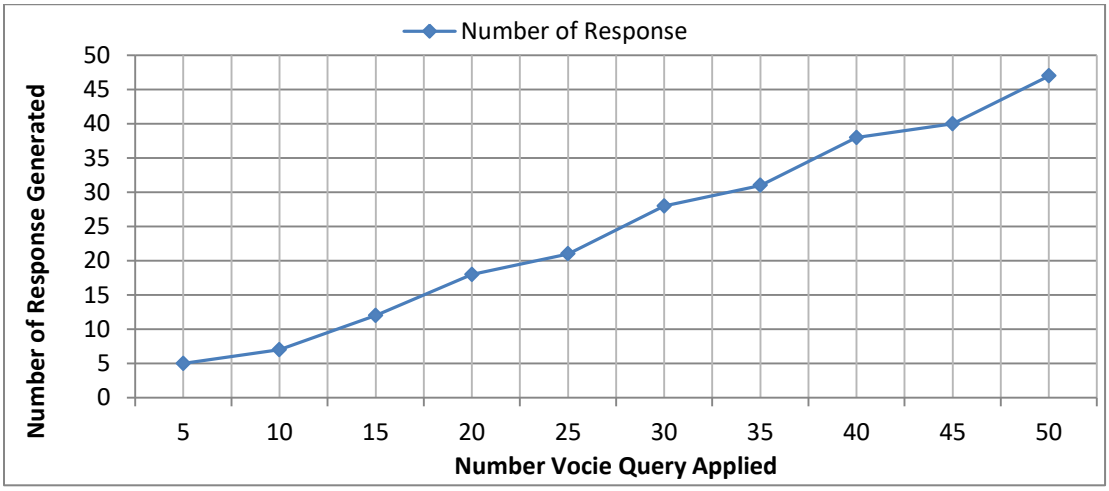


Figure-8: Number Query vs. Number of Response Generated

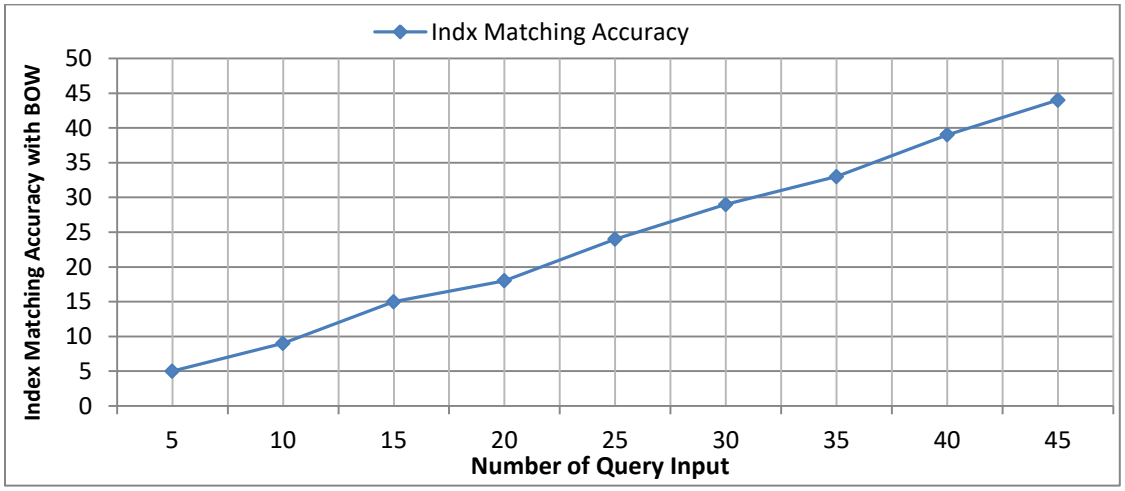


Figure-9: Number of Query vs. Index Matching Accuracy

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

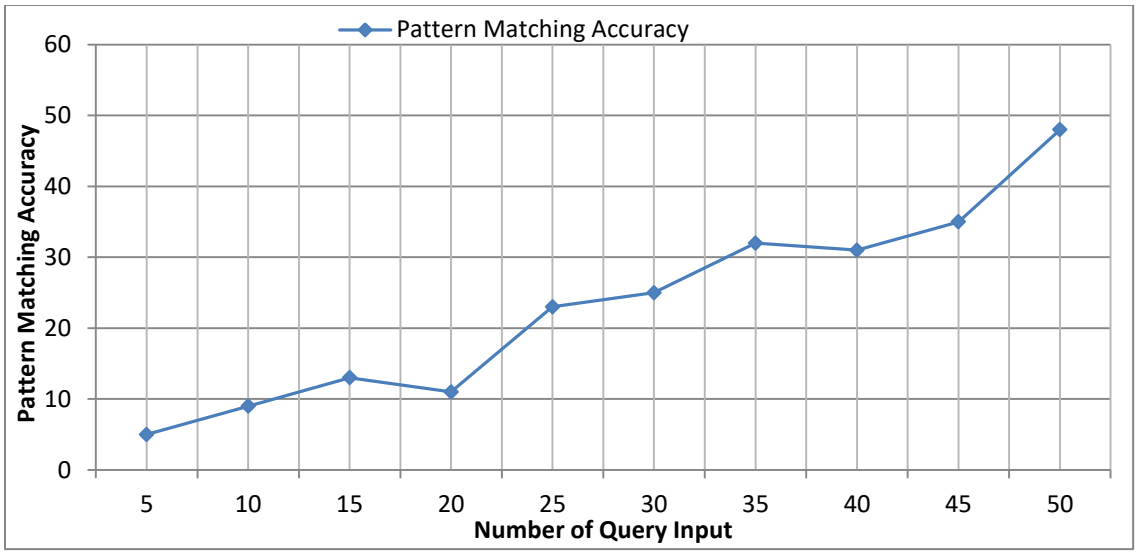


Figure-10: Number of Query vs. Pattern Matching Accuracy

In this paper the number of indices matched and the number of patterns matched is calculated and shown in Figure-9 and Figure-10 respectively. The number of query index matching is proportionally increased, according to the number of query data and accent. The number of pattern matching is up and down in scale due to match pattern and the data available on the DS. In order to evaluate the performance the proposed DRS results are compared with the existing approach.

Performance Analysis

The performance of the DRS is evaluated by comparing the mining accuracy and time complexity with the existing approaches [8]. The proposed DRS and the existing IR system are using the data-dictionary at the back end. The data dictionary size is 100, 120 and 140 in terms of number of words. Figure-10 shows the mining accuracy comparison between proposed DRS and the existing IR [8] system. It is clear that the mining accuracy obtained by the proposed DRS is more than the existing IR. To verify the accuracy and comparability the size of the data dictionary is changed gradually and experimented. In each time of the experiment the mining accuracy is also gradually increased in proposed DRS and it is greater than the existing IR accuracy. Time taken to process the query and response generation and for pattern matching is computed for the proposed DRS and compared with the existing IR system. The time taken by the proposed DRS is lesser than the existing approach time. The experiment is repeated for all the dictionary size 100, 120 and 140, and the time calculated. The calculated time includes the voice processing, BOW index matching and pattern matching time. The complete processing time for one job in the proposed DRS is, time from query word is obtained from voice, compared with the BOW, if exists then it compare with the database. Time taken to process the information retrieved by the proposed and existing is shown in Figure-11. From this figure, it is clear that the time taken by the proposed approach is lesser than the existing approach.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

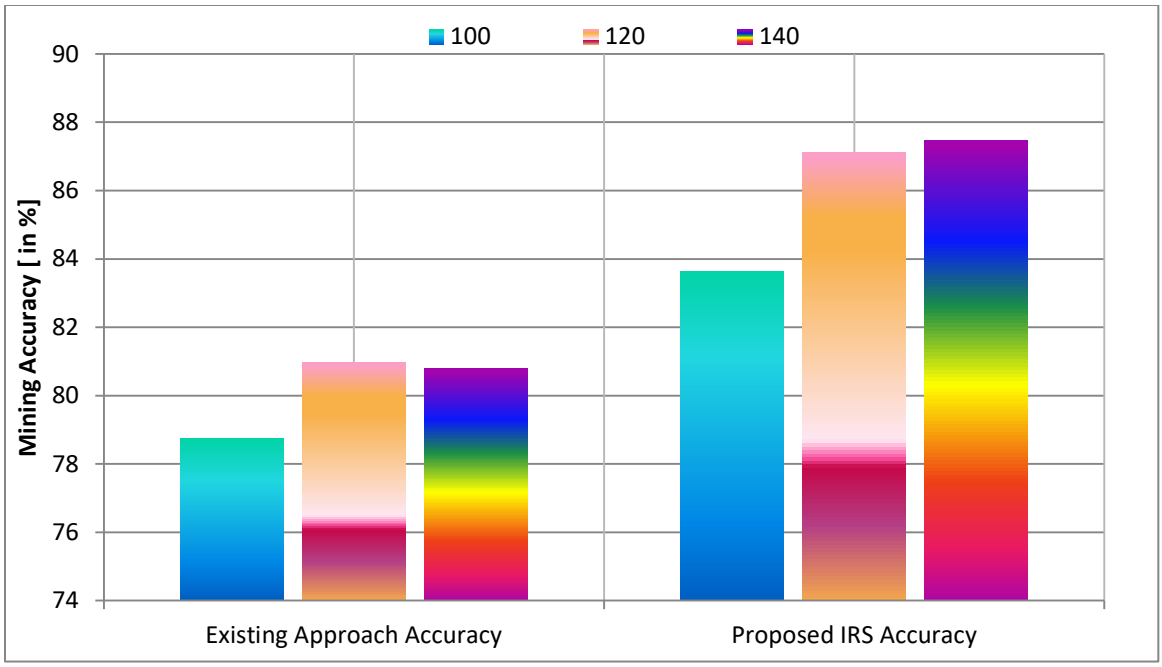


Figure-10: Data Mining Accuracy Comparison between Proposed DRS and Existing Approach

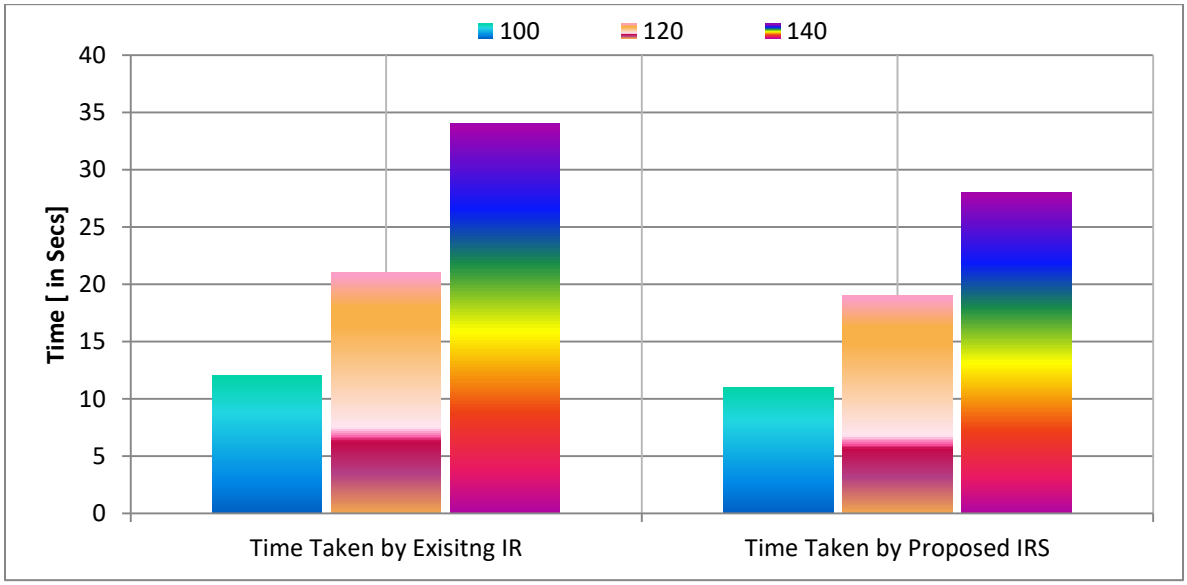


Figure-11: Time Comparison between Proposed DRS and Existing Approach

Run time Efficiency

The efficiency of the proposed DRS is calculated while applying DRS to provide online retrieval and voice reply for large set of database collection. Comparing with the traditional IR approaches, the overhead of proposed voice based DRS comprises four parts: (i). Converting voice to text; (ii). Matching query words in BOW; (iii). Pattern Matching with DB; (iv). Voice based Reply. The previous research off-the-shelf recognition toolkits could already handle the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

entity annotation on queries well with the high accuracy and low latency. By building BOW using the data features, the overhead of index matching and pattern matching process is reduced to do information retrieval. It reduces the time complexity and computational complexity and since this proposed DRS can be extended to large scale data collection, web applications and in wireless network based applications.

Conclusion

The main objective of this paper is to develop a data mining model for physically challenged people using voice. The proposed DRS method uses BOW model in order to retrieve the relevant information from the data. Comparing BOW reduces the computational complexity and searching time. In this paper the proposed DRS handle a smart way of information retrieval approach, which estimate the data availability by comparing the index in order to reduce the time and computational complexity. It can be applied for high – dimensional data entity space. This proposed DRS provides voice to text, text to voice and visual word comparison for improving the efficiency of the information retrieval system. From the results it is clear that this approach is efficient in term of reduced computation complexity, reduced time and it is a special kind of information retrieval system helps to social for physically challenge people like blind and no able to operate keyboard. This voice comparison based authentication can be utilized in various kinds of applications and it is proved.

Reference

[1]. Lakra, Sachin, et al. "Application of fuzzy mathematics to speech-to-text conversion by elimination of paralinguistic content." *arXiv preprint arXiv:1209.4535* (2012).

[2]. KNUTH, D., J. MORRIS, and V. PRATT. 1977. "Fast Pattern Matching in Strings." *SIAM J on Computing*, 6, 323-50.

[3]. BOYER, R., and S. MOORE. 1977. "A Fast String Searching Algorithm." *CACM*, 20, 762-72.

[4]. Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In ICCV, pages 1–8, 2007.

[5]. HHerv´e J´egou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. International Journal of Computer Vision, 87(3):316–336, 2010.

[6]. James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In CVPR, 2007.

[7]. http://www.mysqltutorial.org/mysql-sample-database.aspx.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[8]. Kleber, Florian, Markus Diem, and Robert Sablatnig, "Form classification and retrieval using bag of words with shape features of line structures"-*IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, 2013.

[9]. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In IJCAI, volume 7, pages 2670–2676, 2007.

[10]. M. J. Cafarella, J. Madhavan, and A. Halevy. Web-Scale Extraction of Structured Data. ACM SIGMOD Record, 37(4):55–61, 2009.

[11]. S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.

[12]. T. Lin, P. Pantel, M. Gamon, A. Kannan, and A. Fuxman. Active Objects: Actions for Entity-Centric Search. In *WWW*, pages 589–598, 2012.

[13]. J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *SIGIR*, pages 275–281, 1998.

[14]. J. Pound, P. Mika, and H. Zaragoza. Ad-hoc Object Retrieval in the Web of Data. In *WWW*, pages 771–780, 2010.

[15]. S. E. Robertson and S. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR*, pages 232–241, 1994.

[16]. G. Salton, A. Wong, and C.-S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[17] Kherfi, M.L., Ziou, D. and Bernardi, A. (2004) Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. *ACM Computing Surveys*, **36**, 35-67. http://dx.doi.org/10.1145/1013208.1013210

[18] Datta, R., Joshi, D., Li, J. and Wang, J.Z. (2008) Image Retrieval: Ideas, Influences, and Trends of the NEW Age. *ACM Computing Surveys*, **40**, 1-60.

[19] Yang, M., Kpalma, K. and Ronsin, J. (2010) A Survey of Shape Feature Extraction Techniques. *Pattern Recognition*,1-38.

[20] Penatti Otavio, A.B., Valle, E. and Torres, R.da.S. (2012) Comparative Study of Global Color and Texture Descriptors for Web Image Retrieval. *Int. J. Via.Commun. Image R*, 359-380.

[21] Deselaers, T., Keysers, D. and Ney, H. (2008) Features for Image Retrieval: An Experimental Comparison. *Information Retrieval*, **11**, 77-107.

[22] Mallat, S.G. (1989) A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674-693.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[23] Sarck, J.L., Murtagh, F.D. and Bijaoui, A. (1998) Image Processing and Data Analysis: The Multiscale Approach.

[24] Hill, P., Achim, A. and Bull, D. (2012) The Undecimated Dual Tree Complex Wavelet Transform and Its Application to Bivariate Image Denoising Using a Cauchy Model. 19*th IEEE International Conference on Image Processing* (*ICIP*), 1205-1208. http://dx.doi.org/10.1109/icip.2012.6467082.

[25] Kalra, M. and Ghosh, D. (2012) Image Compression Using Wavelet Based Compressed Sensing and Vector Quantization. *IEEE* 11*th International Conference on Signal Processing* (*ICSP*), **1**, 640-645.

[26] Kokareh, M., Biswas, P.K. and Chatterji, B.N. (2005) Texture Image Retrieval Using New Rotated Complex Wavelet Filters. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*: *Cybernetics*, **35**, 1168-1178.

[27] Balamurugan, V. and Anandha Kumar, P. (2008) An Integrated Color and Texture Feature Based Framework for Content Based Image Retrieval Using 2D Wavelet Transform. *IEEE International Conference on Computing, Communication and Networking*, 1-16. http://dx.doi.org/10.1109/icccnet.2008.4787734

[28] Quellec, G., Lamard, M., Cazuguel, G., Cochener, B. and Roux, C. (2012) Fast Wavelet-Based Image Characterization for Highly Adaptive Image Retrieval. *IEEE Transactions on Image Processing*, **21**, 1613-1623.

Comparative analysis of modern methods and algorithms of cryptographic protection of information

Saleh I. Alomar

AL-Ahliyya amman university Department of Engineering

Saleh A. khawatreh

AL-Ahliyya amman university Department of Engineering

***Abstract*—Information protection problems are topical at the present stage of development of information technologies. Protection of information stored in electronic form, is implemented by cryptographic methods. The article deals with modern symmetric and asymmetric encryption methods. It analyzes advantages and disadvantages of each type of encryption algorithms. Based on comparison results of algorithms, recommendations on the use of algorithms to solve specific problems are provided. The aim of the article is to analyze modern methods and encryption algorithms. When analyzing the strengths and weaknesses of cryptographic methods of protection it is necessary to make a choice of the method of protection on the basis of selected performance criteria, as well as assess the possibility of practical use of the considered cryptographic protection methods for different tasks.**

***Keyword*—**cryptoalgorithm, symmetric algorithm, an asymmetric algorithm, ciphertext.

I. INTRODUCTION

Cryptography, over the ages, has been an art practiced by many who have devised ad hoc techniques to meet some of the information security requirements. The last twenty years have been a period of transition as the discipline moved from an art to a science.

Cryptography is the study of mathematical techniques related to aspects of information security such as confidentiality, data integrity, entity authentication, and data origin authentication.[6]

The constant increase in the volume of confidential information, appearing of new methods and means of unauthorized access to the data leads to the development of information security industry. It is reflected in the creation of new methods and improvement of the existing ones and cryptographic protection algorithms.

The essence of this deficiency is that in the process of breaking any of the known cryptographic systems the cryptanalyst is able to identify the moment their work is successfully completed. This ability stems from the fact that during the cryptographic enciphering of the text the semantic content of the information being protected is, as a rule, transformed into semantically undefined set of alphabet symbols used.[6]

Cryptographic methods and algorithms for protection of information can be divided into:

-symmetric cryptosystem

-asymmetric cryptosystems

Each type of encryption algorithm has its own specific implementation features, advantages and disadvantages that must be taken in dealing with specific problems.

-Symmetric encryption is the oldest and best-known technique. A secret key, which can be a number, a word, or just a string of random letters, is applied to the text of a message to change the content in a particular way. This might be as simple as shifting each letter by a number of places in the alphabet. As long as both sender and recipient know the secret key, they can encrypt and decrypt all messages that use this key.

The problem with secret keys is exchanging them over the Internet or a large network while preventing them from falling into the wrong hands. Anyone who knows the secret key can decrypt the message. One answer is asymmetric encryption, in which there are two related keys--a key

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Symmetric key encryption is a form of cryptosystem in which encryption and decryption are performed using the same key. It is also known as conventional encryption.

Asymmetric encryption is a form of cryptosystem in which encryption and decryption are performed using the different keys – one a public key and one a private key. It is also known as public-key encryption [3].

A Key is a numeric or alpha numeric text or may be a special symbol. The Key is used at the time of encryption takes place on the Plain Text and at the time of decryption takes place on the Cipher Text. The selection of key in Cryptography is very important since the security of encryption algorithm depends directly on it. The strength of the encryption algorithm relies on the secrecy of the key, length of the key, the initialization vector, and how they all work together.

Asymmetric encryption techniques are about 1000 times slower than Symmetric encryption which makes it impractical when trying to encrypt large amounts of data. Also to get the same security strength as symmetric, asymmetric must use a stronger key than symmetric encryption technique. pair. A public key is made freely available to anyone who might want to send you a message. A second, private key is kept secret, so that only you know it.

II. Analysis of symmetric encryption algorithms

Symmetrical encryption algorithm has the key used to encrypt messages that can be obtained from the decryption key and vice versa [2].

In symmetric algorithms, legal user P by means of cipher device Cn turns sequence X = (x1, …, xn), which is called the public information, into the encrypted data Y = Cn(x) (Fig. 1).

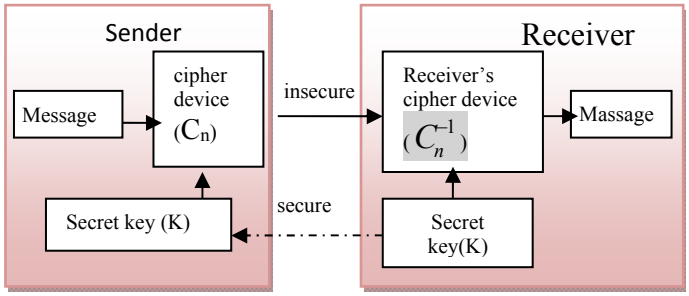


Fig. 1. Structure of the symmetric encryption scheme

The algorithm of the cipher device Cn depends on the parameter K = KX (Key), a known user. Legal users, who possess the information X, perform decryption of information using an algorithm that depends on a parameter K' associated with K. Usually, $K' = K$. In this case, every legal user who originally owns a transformation Cn, and transformation C_n^{-1} - reverse Cn, while the illegal user does not have the key K, which is not fully aware of the conversion Cn and C_n^{-1} [4].

Symmetric cryptosystems are based on the flow and block encryption algorithms. In the flow algorithm, every bit of plaintext is encrypted (and decrypted) by adding module 2 with bit of pseudo-random sequence – cryptographic bit stream, independently of the other bits. Thus, transformation of each intext symbol changes from one symbol to another [5]. Stability of flow encryption algorithm depends on whether the derivative has the property of equal occasional occurrence of the next symbol.

The advantages of streaming algorithms are the high encryption speed, relative simplicity, and the absence error propagation.

The disadvantages are:

– cryptographic bit stream shall not be used more than once (in terms of safety);

– the requirement of operations synchronicity at transmitter and receiver, which is expressed in the transmission timing of a random sequence in front of the message header before its decryption (so-called pseudo-random additional key, which is used to modify the encryption key for improving cryptographic robustness).

A plain text is first partitioned into blocks of equal length for block encryption algorithms, and then is ciphered within each block function depending on the key block into encryption text of the same length [5]. In the case where the length of the plaintext is not aliquant to input block length, multiple encryption algorithm shall be used to supplement operation of the last block of plaintext to the desired length. The essence of the block cipher algorithm is repeatedly applied to the plaintext block of mathematical transformation so as to set a dependency of each bit from the ciphertext and the plaintext key. Block

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

algorithm shall be designed in such a way that the change of even one bit of the plaintext and the key would result in a change of approximately 50% ciphertext bits, while none of the plaintext bit should never be administered directly into the ciphertext [3]. Conversion algorithms based on the data, are divided into a complicated (nonlinear operation) and simple (which are based on mixing), while the first construction provides cryptographic robustness. The most common block encryption algorithms:

1) mode of simple replacement and codebook mode (identical plaintext blocks are encrypted in the same way by the same key);

2) counter mode (initial state is defined by the original range of synchronous communication link, received gamma is processed through block encryption algorithm and then summed in module 2 with the plaintext block);

3) output counter mode (the same synchronous communication and feedback available on the ciphertext, counter mode is performed before the resulting unit will be converted by block encryption algorithm).

The advantages of block encryption algorithms (other than simple replacement mode) are:

Each ciphertext bit depends on all the bytes of the plaintext block and no two plaintext blocks are not represented by the same ciphertext block;

The possibility of application of such algorithms to detect manipulation of the messages made by meddlers.

It uses the fact of error propagation in ciphers and the ability of systems to easily generate a message authentication code.

Disadvantages of block encryption algorithms:

subject to restrictions of cryptanalysis "using the dictionary";

connected with reproduction error (as one error bit in transmission can cause a number of errors in the decrypted text);

development and implementation is more difficult than streaming encryption systems have.

In practice, long messages encryption is applied at inline block algorithms or algorithms with feedback. Repeated alternation of simple permutations and substitutions, managed by a long enough secret key, provides a fairly stable block algorithm with good dispersion and mixing. [3]

The most popular nowadays symmetric encryption algorithms are distinguished: DES, IDEA, GOST 28147-89, Triple, RC2, RC5, BLOWFISH and others.

Each symmetric algorithm is evaluated on the following criteria:

- dimensions of the input and output units;
- key size;
- complexity of data conversion algorithm;
- speed data conversion and cryptoattack resistance.

Stability data rate and conversion was evaluated on 6-level scale (6 - minimum, 1 - maximum [1] (Table 1).

Algorithm	RC5	FEAL	BLOWFISH	ГОСТ 28 147-89	IDEA	DES
Input block size, bit	32,64 or128	64	64	64	64	64
Output block size, bit	32,64 or128	64	64	64	64	64
Key size, bit	from 0 to 2040	64	448	256	128	56
Number of conversion cycles in algorithm	from 0 to 255	From4 to 32	16	32	12	16
Persistence of algorithm	6	4	2	1	5	3
Conversion speed	1	4	3	6	2	5

Table 1, Results of the comparison of symmetric encryption algorithms

Persistence of symmetric encryption algorithms is considered through the following criteria:

- key size;
- complexity of data transformation;
- existence of an algorithm.

In the viewpoint of cryptanalysis, the existence of an algorithm plays an important role. If the algorithm is used for a long time, it becomes an attractive target for cryptanalysts [3] and significant computing resources can be allocated to disclose the encryption algorithm. A DES algorithm can be a famous example of such an algorithm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

According to Table 1, the most resistant to cryptoattacks of enemy is an encryption algorithm GOST 28147-89, but it is considered to be the slowest.

Modern information systems may use symmetric encryption methods in order to prevent unauthorized access to information in the absence of the owner. It can be both an archive encryption of selected files, and automatic encryption of entire logical and physical disks.

Symmetric algorithms are also used to protect data transmitted over open communication channels. [5]

The study of Asymmetric cryptosystems

The essence of the public key or asymmetric cryptosystems of two interrelated keys by a certain rule is generated by each addressee. [4]

The encryption public key scheme is shown in Figure 2.

One key is used for data encryption, the other – for decryption. Each of the correspondents has a key $k = (k_з, k_p)$ consisting of an public key $k_з$ and private key k_p . The open key encryption rule defines Ek, a secret key - decryption rule Dk. These rules are related (for any plaintext X and any ciphertext Y):

$$D_k(E_k(X)) = Y$$

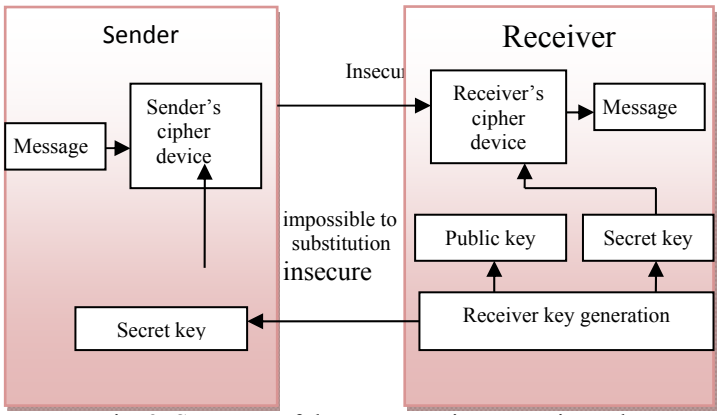


Fig. 2. Structure of the asymmetric encryption scheme

The knowledge of the public key does not allow determining a secret key in a reasonable time (or with reasonable complexity). Let state encryption and decryption rules (on selected key k) of arbitrary correspondent A by EA and DA symbols, respectively. Correspondent B wants to send a private message X to correspondent A, receives a copy EA, calculates the ciphertext Y = EA(X), which directs by communication channel to correspondent A. Correspondent B received message Y, applies DA conversion, receiving plaintext X.

Cryptographic public key systems use irrevocable or unilateral functions that have the following features: given value of X it is relatively easy to calculate the value $f(x)$, but if $y = f(x)$, there is no easy way to calculate the value of X. In other words, it is very difficult to calculate the value of the inverse function $f^{-1}(y)$ [3]. The study of irreversible functions is carried out mainly in three areas: discrete exponentiation; multiplication of prime numbers; combinatorial problems, especially the problem of concluding a portfolio.

Comparison of asymmetric cryptosystems is conducted according to the following criteria: the speed of used algorithms and the mathematical transformation of the information. The data conversion was evaluated on a 5-point scale (1-highest, 5-lowest score). Results of asymmetric encryption comparison techniques are shown in Table 2.

Algorithm	Conversion	Speed
RSA	discrete exponentiation, expansion of factoring	5
Diffie-Hellman	discrete exponentiation	2
El-Gamal	discrete exponentiation	3
Massey Omura	discrete exponentiation	4
Knapsack system	Problem backpack stacking system	1

Table 2.Results of the comparison of asymmetric encryption methods

RSA is considered the most persistent of the existing algorithms, since it is only once failed to disclose RSA cipher for 500-digit key. For these purposes, in 1600 computers of volunteers have been involved in within 5 months of continuous operation [1]. It should be noted that using the RSA system with keys 512-1024 bits is practically impossible to break ciphers. However, RSA system operates in thousand times slower than DES algorithm, and requires that the keys to be approximately 10 times longer.

While it is clear that the use of public key systems can be limited by challenge key exchange, followed by their use in symmetric cryptography that is the use of so-called hybrid systems [4]. The results of the comparison of classical cryptographic algorithm DES and cryptographic algorithm RSA with public key are shown in Table 3.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Characteristic	DES	RSA
Speed	Fast	Slow
Function used	Permutation and substitution	Involution
Length of the key	56 bit	300…600 bit
Least expensive cryptanalysis	Iterate over the key space	Module decomposition
Temporary costs on cryptanalysis	Centuries	Depends on the key length
Key generation time	Millisecond	Tens of seconds
Type of key	Symmetric	Asymmetric

Table 3 Results of comparing DES and RSA algorithms

When analyzing the strengths and weaknesses of symmetric and asymmetric systems, it is determined that the asymmetrical encryption systems provide a significantly lower encryption rate than symmetrical, that is why they are usually used not only to encrypt messages, but as encryption of keys exchanged between correspondents, which are then used in symmetric systems.

The main advantage of public key cryptosystems is their potentially high safety: there is no need to transfer or disclose to anyone the value of the secret key, to make sure of their reliability. In symmetric cryptosystems, there is the risk of disclosure of the secret key during the transmission.

However, the algorithms that base public key cryptosystems have the following disadvantages:

- Generation of new private and public keys based on the new generation of large prime numbers and primality testing takes a lot of device time;
- encryption and decryption processes are related the construction of the power of a multi-valued number, are rather cumbersome.

Therefore, the speed of public key cryptosystems is usually hundreds times or even more less than the speed of symmetric sector key cryptosystems.

Asymmetric encryption algorithms are used to solve many problems: user authentication and message, generation of session keys in information systems, systems for identification “friend-or-foe”.

III. Conclusions

- The studies of modern methods and algorithms for cryptographic protection of information from unauthorized access can conclude that modern information systems for the encryption of transmitted messages use symmetric encryption algorithms. Asymmetric algorithms, because of their large computational complexity, are used for the generation and propagation of session keys.
- The combined use of symmetric and asymmetric encryption allows eliminating the main drawbacks of both methods.
- The combined method of encryption keeps the advantages of high security provided by asymmetric cryptosystems with a public key, and the advantages of high speed operation, inherent in symmetric cryptosystems with a secret key. The proposed approach allows choosing the method of protection based on the selected performance criteria, as well as assessing the possibility of practical use of the considered cryptographic methods of protection.

References

[1]. Burnett S., Payne S. RSA Security's Official Guide to Cryptography. – M.: Bean, 2002. – 382 p.

[2]. Rostovtsev A.G., Mahovenko E.B. Theoretical cryptography. – M.: Publ. Professional, 2005. – 490 p.

[3]. Salomaa A. Public key cryptography. M.: Mir, 1995. — 318 p.

[4]. Schneier B. Applied Cryptography: Protocols, Algorithms, Source codes in C language. — M.: Triumph, 2002. — 816 p.

[5]. Blahut R.E. Cryptography and Secure Communication. – Cambridge University Press, 2014. – 608 p.

[6]. Hazem Hatamleh (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 10, 2016

SQS: An Ontology Based Framework for Evaluating Service Oriented Software Applications: A case study of E-Governance Services

Ezekiel U. Okike
Department of Computer Science,
University of Botswana, Gaborone, Botswana
euokike@gmail.com

***Abstract*—E-governance is about enabling good governance through the use of modern Information and Communication Technology. As a service based ICT platform, the main challenge is efficient and effective evaluation framework. In this paper, an ontology based framework for evaluating e-government software applications is proposed. The proposed framework uses a three stage model: standardization, quality, and service stages (SQS). The model provides effective and dependable evaluation of e–governance from users and stakeholders' perspectives.**

Keywords-ontology; e-governance; framework; evaluation model .

I. INTRODUCTION

A Ontology is a formal approach to specifying a concept and its representation of a domain [1]. The concept being represented is explicitly described using formalisms or other appropriate representation that provides a description of the concepts and the relations between them as well as its technological components [2,3].

Using ontologies computational models are created for automated reasoning in Artificial intelligence [4]; classes, relations, functions and objects are defined in Object Oriented systems [5]; common understanding of objects are shared; knowledge reuse, and explicit assumption are enabled; and domains are separated and analyzed [6,4]. Furthermore, ontologies are used in classifying object based on scope or domain granularity, taxonomy construction direction, and the type of data sources [7]. Figure1 shows the various levels of ontology classification: the base level (application level), the intermediate level (domain oriented and task oriented), and the top level.

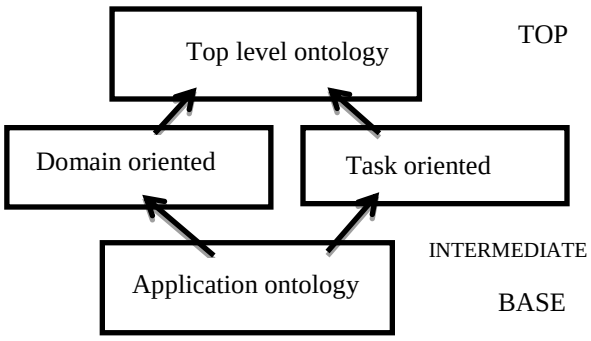


Fig. 1. Ontology classification
Source: Adapted from Antonio [8].

In this model, the top-level ontology has some concepts that have general agreements or stable standards. Domain ontology has concepts that define the main focus of interest on the domain. Task ontology deals with sub-concepts that are needed to solve problems on the main domain and the application ontology deals with concepts that exercise the fastest rate of exchanging data [8].

A. Problem Statement

The search for effective and efficient evaluation model of e-government services continues as current evaluation option are still evolving especially in developing countries. This is so because in the developing countries the concept of e-governance is still poorly implemented and lacks appropriate standards. Hence, the need for efficient evaluation framework cannot be over emphasized.

B. Objectives

The main objective of this paper is to study existing e-government models and hence propose appropriate framework for evaluating e-government software services.

This rest of this paper counts of three sections. Section 2 is a review of the literature on e-government concept and how to build an ontology for e-government. Section 3 looks at

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ontology based e-government models. Section 4 proposes an evaluation model for e-government, while section 5 is conclusion and future work.

II. LITERATURE REVIEW

A. The Concept of E-governance

E-governance is about enabling good governance through the use of modern Information and Communication Technology (ICT). The concept (also known as Digital governance) implies the growing use of ICT as a catalyst for the formation of knowledge societies where people have more access to relevant information as participants in their own governance and development. According Nath [9], "Knowledge networks function on the underlying principle that access to information is empowering and strategic use of information by citizens could become the key to popular and meaningful governance". This assertion is premised on the knowledge networking model shown in figure 2.

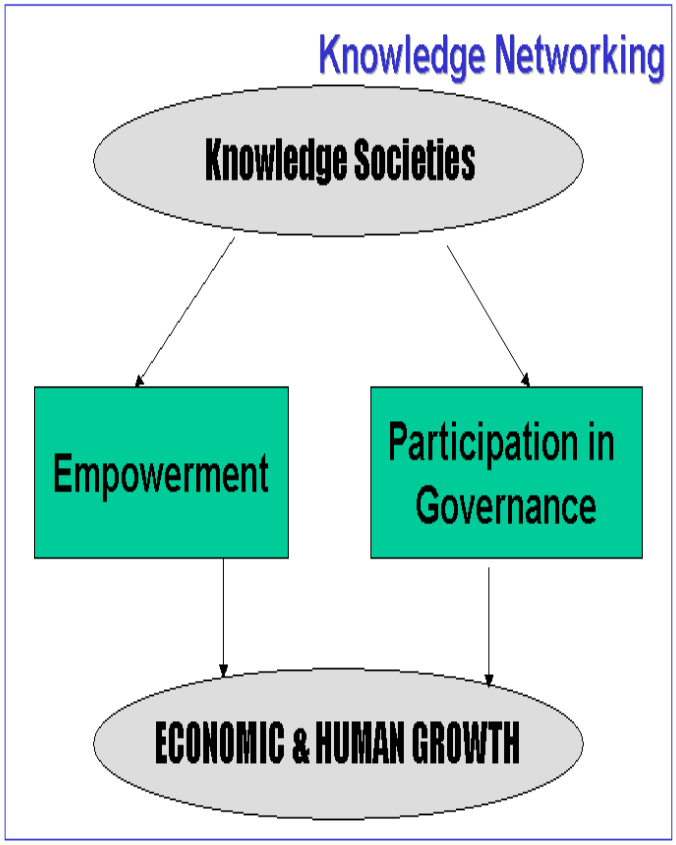


Figure 2: Knowledge Networking through ICT empowerment (Source: Nath, V [9])

Although, e-governance (digital governance) is still evolving in developing countries, there are five generic models in use [9]. These include Broadcasting/wider dissemination model, Critical Flow model, Comparative Analysis model, Mobilization and Lobbying model, and Interactive–service model. The underlying principle, applications and organization of each model is summarised as follows:

i) Broadcasting/Wider-Dissemination Model – This model aims to disseminate information for better governance through the use of ICT. This helps the citizenry understand governance so that they are able to make informed decisions.

ii) Critical Flow model - This model aims to channel information of critical value to targeted audience through the use of ICT. Using ICT such information is disseminated timely irrespective of distance.

iii) Comparative Analysis model- This model aims to explore information available in the public or private domain, and compares that with already known information for strategic purposes. Therefore, new and assimilated information are used as benchmark for governmental advocacy and policies

iv) Mobilisation/Lobbying model- This is a digital governance model often used by civil society organizations in order to make their influences and impacts known through virtual communities.

v) Interactive-service model – this model aims to offer government services to the citizens using interactive ICT channels such as e-voting, e-tax, e-procurement e.t.c

B. Building an E-governance Ontology

Several approaches could be followed to build an ontology based e-governance. One could either use the bottom–up approach, top-down approach or the middle out approach (Catherine Roussey et al, 2011).

Bottom-Up approach: defines first the most general concept of the entity in use then goes towards the most specific aspects.

Top-Down approach*:* defines first the most specific concepts then goes towards the most general aspects.

Middle-Out approach*:* defines the concepts from the central area towards the general and / or specific concepts. Therefore, an e-government ontology may be defined following these principles.

According to Roussey [7], ontologies could also be described according to sources used to get the knowledge. The knowledge could either be based on:

Text*:* Unstructured data given to a computer system for processing.

Thesaurus*:* forming concepts from words or linguistic relations to build ontology.

Relational Database*:* structured and accurate software storages used to build ontologies from.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

UML Diagrams*:* using formal described UML classes to define concepts to build ontologies

In addition, an e-governance ontology can also be defined using the Enterprise Ontology Modelling Process (EOMP) identified by Uschold and Gruninger [10]. Using this approach requires the following:

i) Identify Purpose and Scope: which deals with main reason why the ontology is being built

ii) Building the ontology**:** segmented into three steps

(a) Ontology capture*:* deals with identifying the key concepts and relationships in the domain of interest.

(b) Ontology coding*:* deals with representation of the knowledge using a formal language for the ontology.

(c) Integrating existing ontologies*:* incorporates the both coding and capturing process with logic of how to use the ontology.

(d) Evaluation: gives a technical judgment on the ontology

(e) Documentation: Stating the guidelines for each purpose

Furthermore, ontology development process could be done following the IEEE standard for developing Software Life Cycle Process [11].

III. ONTOLOGY BASED E-GOVRNANCE EVALUATION MODEL

E-governance is a software based online/web based service. Hence, some of the principles of measurement in software are very useful in evaluating e-governance structures.

A. Ontologies in Software measurement

Generally, measurement is a mapping from the empirical world to the formal, relational world**.** Consequently, a measure is the number or symbol assigned to an entity by this mapping in order to characterize an attribute [12].

Theoretically, Measurement Theory (MT) species the rules for developing and reasoning about all kinds of measurement. As explained in [14], rule based approach is common in the sciences such as Chemistry, Physics and Mathematics. In Mathematics, Mathematicians learned about the world by defining axioms for a geometry. Hence, by combining axioms and using their result to support or refute their observations, they expanded their understanding and the set of rules that govern the behavior of objects.

In any software measurement activity the entities and attributes to be measured must be clearly identified and specified.

In software measurement, three software activities are involved namely:

i) Processes – collections of software related activities

ii) Products - artefacts, deliverables or documents resulting from process activities

iii) Resources – entities required by a process activity

Software artefacts have 2 essential types of attributes namely internal and external attributes.

Internal attributes are measured in terms of the product itself. Essentially, internal attributes are code based measure of software quality attributes such as class cohesion, class coupling, control structures, algorithms, data structures, and nesting level[13].

External attributes are measured in terms of how the software product, process or resource relate to the environment of operation. The measures are aimed at evaluating the software from the users perspectives in terms of its usability, reliability, efficiency, reusability, maintainability, portability, and testability e,tc. Figure 3 below shows the standard ISO/EC 9128 evaluation guide based on external software attributes. This guide is a useful ontology based model for all aspects of internal and external software quality measures.

External attributes (figure 2) are measured in terms of how the software product, process or resource relate to the environment of operation. The measures are aimed at evaluating the software from the users perspectives in terms of its usability, reliability, efficiency, reusability, maintainability, portability, testability e.tc. ISO 9126 [15] proposed a standard which species six areas of importance, i.e. quality factors, for measuring external software attributes. These include functionality, reliability, efficiency, maintainability, portability, and usability. This model was has since evolved into the ISO/EC 9128 [16] software product evaluation standard as shown in figure 2.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

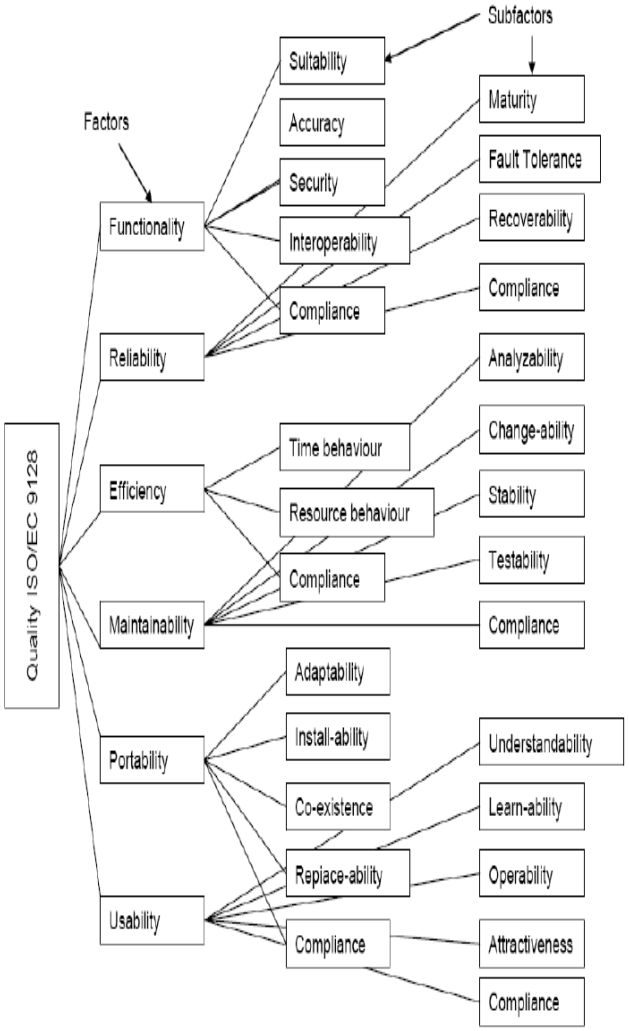


Fig. 2: ISO/EC 9128: Software Product Evaluation: Quality Characteristics and Guidelines for their Use.

This model has evolved into ISO/EC 25010. A detailed review of software quality models for the evaluation of software products is presented in Miguel, Mauricio, and Rodriguez [17]. However, in this paper, although all the models are useful, the ISO/EC 9128 standard is used. By integrating standard e-governance model, and the ISO/EC 9128 or ISO/EC 25010 this paper proposes an evaluation framework for e-governance as described below.

IV. SQS:AN ONTOLOGY BASED FRAMEWORK FOR EVALUATING E-GOVERNANCE.

SQS is an acronym for Standards, Quality and Service. Hence, the proposed e-governance framework is focused on the following aspects:

i) Standards. Any e-governance evaluation should begin by ascertaining if the e-governance in place is modelled after acceptable e-governance standard such the one defined by
   a) The broadcasting model
   b) The critical flow model
   c) The organisation/project based model
   d) The comparative analysis model
   e) The mobilization and lobbying model
   f) The interactive service model

The key question to answer is "Does existing e-governance follow acceptable standard?" i.e. does it take care of the items "a-f" in its implementation ?.

ii) Quality (Quality of Service QoS). The QoS of a Service Oriented Software Initiative (SOSI) such as e-governance will be better evaluated using both the internal and the external software quality attributes such as the standard ISO/EC 9128-ISO/EC 25010 software product evaluation quality characteristics. This can be done by designing appropriate questionnaires which capture all desired external attributes for the users of the e-governance service. By analysing collected feedback, and interpreting results, a good evaluation of any e-governance service may be obtained in terms of its QoS based on the factors identified in figure 2.

iii) Service Delivery
A Service Delivery Framework (SDF) is a set of principles, standards, policies and constraints to be used to guide the design, development, deployment, operations and retirement of services delivered by a service provider with a view to offering consistent service experience to a specific user community in a specific community. The important question to answer in evaluating an e-governance is "Is there any Service Delivery (SDF) model in place? This implies ascertaining that principles, policies, standards and constraints in respect the existing e-governance are in place. If these are in place, the next question to answer is "Is service delivered ?" "By what indicators?" Measurable indicators of service delivered could be achieved by :
   a) Specifying expected output indicators
   b) Ascertaining service effectiveness
   c) Ascertaining user satisfaction
   d) Ascertaining service availability
   e) Ascertaining service functionality
   f) Ascertaining service reliability
   g) Ascertaining service measurability
   h) Ascertaining service accountability

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

i) Ascertaining service manageability e.t.c

Outcomes are the end result that the government wishes to achieve through its e-governance initiative, and in particular with reference to how the rural populace benefit from the e-governance service. Indicators assess the impact of the program output on the desired outcomes that government want to achieve in the e-governance initiative.

A. *Measuring Service delivery*

The following steps are necessary in order to measure service delivery:

a) Clarify service delivery and performance measurement tools
b) Specify appropriate measureable objectives and output
c) Develop robust output measures and indicators.

B. *Relationship between internal and external Attributes*

Internal software attributes are code level measures of the quality of the underlying codes of the software. Some code level measure include cohesion, coupling, lines of code, cyclomatic complexity, Depth of inheritance

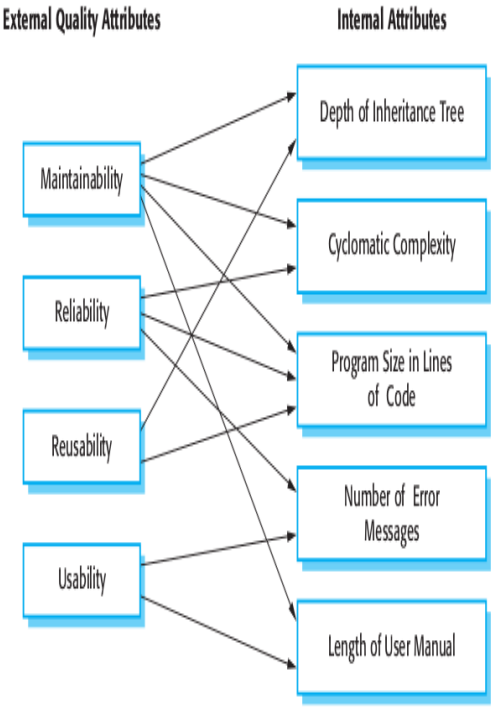


Figure 4. Relationship between internal and external quality attributes

For evaluating e-governance, using internal attributes of the software are not recommended, but using external attributes are highly recommended. This is because rural users are the object of measuring the success of e-governance initiatives.

V. CONCLUSION

As a service oriented software platform, e-government success hinges on service delivery. Successful service delivery models are based on appropriate standards, and policies which are also part of the software implementation.

In this paper, a three stage model for evaluating e-governance has been proposed. The stages in this evaluation model include standardization, quality and service (SQS) . The future direction of research on this paper will focus on empirical studies based on the SQS framework.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] Kang Ye etal. (2009). Ontologies for crisis contagion management in financial institutions. *Journal of Information Science, 35 (5)*, 548–562.

[2] Silvonen, P. (2002, October 21). *Ontologies and Knowledge Base*. Retrieved October 22, 2015, from http://www.ling.helsinki.fi: http://www.ling.helsinki.fi/~stviitan/documents/Ontologies_and_KB/ontology.html

[3] Sunitha Abburu and G Suresh Babu. (2013). A Framework for Ontology Based Knowledge Management. *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-3*, 21-25.

[4] Gruber, T. (2009). Ontology. In L. L. Özsu (Ed.), *Encyclopedia of Database Systems.* Springer-Verlag.

[5] Gruber, T. R (1993). A Translation Approach to Probable Ontology Specification. Knowledge Acquisition, 5(2):199-220

[6] Kabilan, V. (2007). *Ontology for Information Systems (O4IS) Design Methodology: Conceptualizing, Designing and Representing Domain Ontologies.* The Royal Institute of Technology: KTH information and Communication Technology .

[7] Catherine Roussey et al. (2011). An Introduction to Ontologies and Ontology Engineering. In F. G. al, *Ontologies In Urban Development Projects* (p. 241). London: Springer-Verlag.

[8] Antonio Zilli et al. (2009). *Semantic Knowledge Management: An Ontology-Based Framework: An Ontology...* New York: Information Science Reference

[9] Nath, V (2007). Digital Governance Models: Towards Empowerment and Good Governance in Developing Countries in Rafael Capurro, Johannes Frühbauer, Thomas Hausmanninger (eds). Localizing the Internet. Ethical Aspects in Intercultural Perspective. Munich . ISBN 978-3-7705-4200-0

[10] Mike Uschold and Micheal Gruninger. (1996). Ontologies: Principles, Mthods and Applications. *Knowledge Engineering Review Volume 11 No 2* , 1-69.

[11] Mohammad Nazir Ahmad et al. (2012). *Ontology-Based Applications for Enterprise Systems and Knowledge Management.* United States of America: IGI Global.

[12 Fenton, Norman and Pfleedger, S. L (1997). Software Metrics: A Rigorous and Pratctical Approach, 2nd ed, Boston, M.A:PSW Publishing

[13] Ezekiel U. Okike, and Tebo Leburu (2017). Axiomatic and Measurement Theory based Framework for Selecting Success Factors in Software Project Management. International Journal of Computer Science and Network Security, Vol. 17 No. 12 , pp.

[14] Okike, Ezekiel U (2007). Measuring Class cohesion in Object oriented Systems using Chidember and Kemerer Metric Suite and Java as case study. PhD Thesis Department of Computer Science, University of Ibadan. Unpublished

[15]ISO/EC9126.http:/www.bth.se/com/besq.nsf/(webfiles)/…/$FILE/chapter_1.pdf

[16] ISO/EC9128http:/www.bth.se/com/besq.nsf/(webfiles)/…/$FILE/chapter_1.pdf

[17] Miguel, Jose P; Mauricio, D; and Rodriguez, G (2014). A review of Software Quality Models for the Evaluation of Software Products. International Journal of Software Engineering & Applications (IJESA), Vol. 5, No. 6, pp. 31-53.

AUTHORS PROFILE

Dr Ezekiel Uzor Okike, MIEEE, MACM, MAIS is a Senior Lecturer in the Department of Computer Science, University of Botswana, Gaborone. He holds a PhD in Computer Science (2007), a Master of Information Science (M.Inf,Sc), 1994 and a B.Sc Computer Science, (1992) all from the University of Ibadan, Nigeria. He is a member of IEEE, ACM, and AIS. Presently he is the cluster chair of the Information Systems Cluster, Department of Computer Science, University of Botswana. He has published many papers in international journals and attended many international conferences where he presented papers. His research interests are in Software Engineering, Software Measurement and Models, Software Architectures, Information Systems, Cyber Security, and High Performance computing.

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2017-2018
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/